A clean slate

Nature is pleased to name physicist Steven Chu, Nobel laureate and the US Secretary of Energy, as its Newsmaker of the Year.

Steven Chu made his name — and earned his Nobel prize — by developing an ingenious laser technique for capturing and studying atoms. He is an extraordinary experimentalist who loves the challenges of the lab. But five years ago, he embraced a much bigger challenge when he took the helm at Lawrence Berkeley National Laboratory in California and dedicated it to clean-energy research. Chu was sworn in as secretary of the US Department of Energy this January, and is now charged with transforming the way the world's largest economy powers itself. That is why *Nature* has selected Chu as its Newsmaker of the Year (see page 978).

Chu has already had a significant impact. From his position near the top of President Barack Obama's administration, he has helped make the case that the United States must commit to reducing its greenhouse-gas emissions, not only to save the planet but also to ensure that the country will be able to compete with China, India and Europe in the emerging green economy.

And he has started to commit the energy department to doing more high-risk research on clean energy. This is the proper role of governmental research programmes: to develop promising technologies that are too far from the market to draw the support of industry, which has to focus on near-term profits. Nations cannot simply regulate themselves out of the global-warming problem; pledges to cut greenhouse emissions have to be made real by the scientists and engineers who develop cheaper, cleaner and more efficient ways to produce and distribute energy.

Chu has accordingly made it his mission to be a public cheerleader for clean-energy research. As he travels around the United States, he trumpets the need for the best minds of this generation to tackle perhaps the largest problem that society now faces. He has already recruited some top scientists to leave their chosen fields of research and — just as he did — focus on the energy issue.

Such commitment to public service is a long and honourable tradition among scientists. Thus, *Nature*'s selection of Chu as Newsmaker is also a tribute to the legions of scientists around the world who step away from research, sometimes for life, to tackle the larger problems of society.

There have been times when Chu has struggled during his first year of tenure. Scientists who take leading roles in government are often selected because they, like Chu, are outstanding researchers. But the skills that matter most in a laboratory are not enough to ensure success in the arena of public policy and politics, where building consensus and finding acceptable compromises are key. As Chu is learning, researchers entering government need to adapt and learn from political leaders.

> **"Steven Chu has made it his mission to be a public cheerleader for clean-energy research."**

At the same time, scientists can help politicians to think more scientifically — encouraging them to examine the results of research carefully rather than just follow ideology and political expediency. Chu has taken it upon himself to meet with legislators, particularly those who do not believe that global warming is a problem. He walks them through the science of climate change and the consequences of ignoring it, all with the aim of helping to win passage of a climate bill that will clean up the US energy sector.

Although it would be naive to think that one scientist can move politicians with long-standing opposition to climate legislation, Chu's efforts may help tilt the balance towards the passage of a bill and the eventual ratification of an international climate treaty. Those who stand on the side of logic and data should vigorously support his efforts. ■

After Copenhagen

The agreement reached last week lends fresh urgency to challenges in science and communication.

It is easy to feel disappointed by the accord brokered last week by US President Barack Obama at the United Nations Framework Convention on Climate Change in Copenhagen. The document's broad outlines do not constitute a treaty, nor is it even clear whether it should technically be called a global agreement. Crafted principally by a handful of nations — the United States, China, India, Brazil and South Africa — the accord was presented as a take-it-or-leave-it fait accompli to representatives of the nearly 200 other nations in attendance, few of whom had been consulted. The resulting protests nearly led to the convention's collapse (see page 966).

The accord provides a framework for capturing the national climate commitments already on record, but has no mechanism to enforce them. Nor does it offer any global emissions targets. Even if fully implemented, the accord would allow greenhouse-gas emissions to continue rising beyond 2020, and would put the world on a course towards a warming of nearly 4 °C by 2100. It would not halt the acidification of the world's oceans or the melting of its ice. It is so weak that many critics worry that it will actually undermine progress by failing to send a clear signal to markets, investors and governments about the need to reduce greenhouse-gas emissions.

Nonetheless, for the first time, all of the world's largest greenhouse-gas emitters have signed up to a framework for cooperation on the biggest challenge of our time. For all of its shortcomings, the accord is an important step forwards.

In many ways, the accord represents a snapshot of humanity at a crucial juncture in history. Tensions between rich and poor, north and

south, past and future, all came to the fore during the Copenhagen conference, and right up to the moment Obama announced the accord, it was unclear whether the meeting would produce anything. Even then, the agreement was so controversial that the convention was unable to adopt it as a formal decision; instead it was recognized as a proposal that countries can sign up to if they so choose.

In fairness, formal adoption was blocked by just a handful of countries — which was all it took under the UN consensus rules that governed the conference. Those rules, which make it extraordinarily difficult to get anything done through the UN framework, are perhaps the strongest argument for the more bottom-up approach epitomized by the Obama deal. Instead of waiting for a top-down, global consensus, it may be better to build on the international momentum that has been achieved by individual nations pursuing their own goals on climate.

Such individual efforts can reinforce one another, even without explicit global coordination. Consider the climate bill that is pending in the US Senate. Members of that body are wary of committing to caps on greenhouse-gas emissions if they think that developing countries can continue to emit greenhouse gases at will — a perception that doomed Senate ratification of the 1997 Kyoto Protocol. This time, all of the major emerging economies prepared for Copenhagen by assessing what domestic greenhouse-gas cuts they were prepared to make. This is crucial for any meaningful action on climate, as the developing world will be responsible for virtually all of the growth in greenhouse-gas emissions in the decades to come. It also means that these countries are more likely to meet their commitments, because the goals are not just numbers imposed from above by a UN-wide consensus. Under the accord, moreover, those commitments will be public knowledge, which should make it easier for the international community to review each nation's performance — a requirement that Obama explicitly sought in Copenhagen. That, in turn, should help his administration in the monumental task of pushing the climate bill through Congress next year.

Investments in the future
The developing countries, meanwhile, got the Copenhagen accord's promise of some US$30 billion during the next three years and $100 billion annually by 2020 to help them cope with global warming. Some of that money will go towards forest conservation, helping to curb the significant releases of carbon that result from deforestation (see *Nature* **461,** 1048–1052; 2009). But the money could also be used to help poor nations adapt to a changing climate, work on sustainable development plans or pay for low-carbon energy technologies. This promise represents a significant transfer of wealth from rich nations to poor, and may be only a down payment: some advocates have argued that the $100-billion commitment must be tripled to get the job done. Many politicians in the rich nations have rightly come to view this money not as a donation, but as an investment; everyone will pay a far larger price down the road if poor countries follow the conventional carbon-intensive pathway to development.

In an ideal world, negotiators would have come to the table with such commitments and then used them as a foundation for an even stronger agreement. But that kind of negotiating requires trust,

> "Scientists need to develop a better understanding of the full suite of physical processes affecting the planet."

which proved to be in short supply in Copenhagen. For now, the world will have to live with a collection of national commitments that can, over time, be built into something stronger.

The next opportunity will be during the first review of implementation and commitments under the accord, which is scheduled to be finished by 2015, one year after the completion of the fifth assessment of the Intergovernmental Panel on Climate Change (IPCC). The IPCC's alarming fourth assessment in 2007 set the stage for Copenhagen, and the world mobilized as a result. The fifth assessment will be just as important.

But key messages for citizens and their leaders have emerged from the continual observations of the state of the planet since the fourth assessment. It is essential that the vigour of such research be sustained even in the face of tight budgets and an anaemic world economy. Scientists need to develop a better understanding of the full suite of physical processes affecting the planet, from cloud formation to feedbacks that affect the way the climate responds to greenhouse gases and other perturbations, whether natural or caused by humans. They particularly need the computational resources to improve regional climate forecasting. Although there is good agreement about how climate will change on a large scale, there is considerable uncertainty over changes at the local and regional level, which is the kind of detail required by governments for planning purposes.

Improved understanding
Researchers have also made significant progress in assessing and monitoring carbon in forests, but need to do more to develop a holistic view of carbon trends across farms, cities, wetlands, oceans and every other part of Earth. Also crucial is a renewed emphasis on agricultural research — until recently a comparatively neglected discipline, but one that is essential to counter the effects of global warming, feed a growing population and support efforts to reduce deforestation, of which agriculture is a primary driver. The social sciences, which are rightly receiving more emphasis in the IPCC's fifth assessment exercise than they have in previous rounds, are likewise important: a better understanding of people, communities and society should be central to devising solutions that realistically can be implemented.

A crucial challenge for researchers over the next few years will be to ensure that public communication of the results of climate-change research — whether formally or in the blogosphere — is robust in every sense: making the results comprehensible, and even vivid, yet rigorous; doing full justice to the uncertainties; maintaining such standards in the face of misinformation and propaganda; and responding promptly to the unexpected.

But governments have the biggest challenge of all. In the face of major threats to their countries' long-term futures, they need to commit to an international treaty. They also need to sustain the bottom-up motivation for change. They must empower their citizens with information: not only about what science is saying, but also about citizens' own contributions to the problem. New energy technologies will be required, but the quickest returns on investment in mitigating climate change and its effects can be found in energy efficiency and incentivizing behavioural change. ■

RESEARCH HIGHLIGHTS 2009

As the year draws to an end, *Nature*'s editors look back on some of their favourite papers published elsewhere this year.

ATMOSPHERIC SCIENCE
Stealth ozone destroyer

Science **326,** 123–125 (2009)

Thanks to the Montreal Protocol, emissions of chlorofluorocarbons (CFCs) have dropped, resulting in early signs of recovery of the ozone layer. However, the battle is not yet won. Researchers report that nitrous oxide (N_2O) — a gas that behaves similarly to CFCs in triggering global ozone destruction, but is not regulated by the protocol — is now the most significant ozone-depleter being emitted.

A. R. Ravishankara and his colleagues at the National Oceanic and Atmospheric Administration in Boulder, Colorado, calculated the 'ozone depletion potential' of N_2O and used it to weight the current emissions level. They found that the 2008 weighted emissions level for N_2O was around double that of the next most important ozone-depleting gas.

They also noted that nitrous oxide's global-warming potential is second only to methane among non-carbon dioxide gases in terms of its contribution to warming.

STEM CELLS
Protein reprogramming

Cell Stem Cell **4,** 381–384; 472–476 (2009)

In 2006, researchers made cultured skin cells behave like embryonic stem cells by adding a handful of genes. But the method by which they inserted the genes — using viruses — can turn cells cancerous.

This year, two groups accomplished the same feat by delivering just the protein products of these genes into cells. Sheng Ding at the Scripps Research Institute in La Jolla, California, and his colleagues converted mouse embryonic fibroblasts into embryonic-like stem cells by using the bacterium *Escherichia coli* to engineer modified versions of the proteins that could cross the cellular and nuclear membranes. The cells (pictured green, right) were able to incorporate into early mouse embryos.

Meanwhile, Robert Lanza of Stem Cell and Regenerative Medicine International in Worcester, Massachusetts, Kwang-Soo Kim of Harvard Medical School in Boston, Massachusetts, and their colleagues reprogrammed human fibroblasts. They used a human cell line to generate the proteins, fused with a cell-penetrating peptide.

These papers are just two from this year that improved methods of stem-cell reprogramming.

NEUROSCIENCE
Signal source questioned

Proc. Natl Acad. Sci. USA **106,** 18390–18395 (2009)

Functional magnetic resonance imaging (fMRI) maps brain activity by measuring changes in local blood flow and oxygen levels. Neuroscientists have long thought that most changes in these signals were due to neurons becoming active and requiring more oxygen.

A study in October by Aniruddha Das and his colleagues at Columbia University in New York challenged this idea. They used intrinsic signal optical imaging, a technique similar to fMRI, to measure changes in blood volume and oxygenation in the brains of two macaques while the monkeys performed a visual task. The researchers simultaneously recorded local neuronal activity.

They showed that the initial part of the imaging signal is marked by changes in blood volume, not in blood oxygen levels. The authors concluded that blood volume is more tightly linked to neuronal activity and is thus a better measure to use in brain imaging. They say that boosting blood volume is the brain's way of 'anticipating' neuronal activity to meet the cells' impending need for more oxygen.
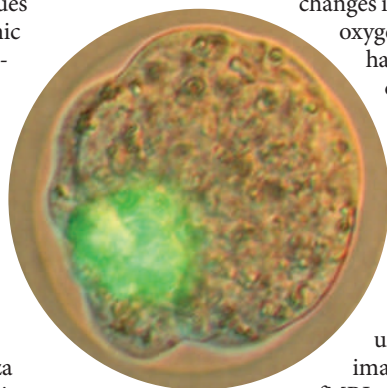
Not from chimps

Science **326,** 75-86 (2009)

In October, after 17 years of investigation, researchers reported detailed descriptions of the oldest hominid skeleton yet found. The fairly complete 4.4-million-year-old female *Ardipithecus ramidus* fossil — known as Ardi — was discovered in Ethiopia.

Analysis of the remains by Tim White at the University of California, Berkeley, and a large team of collaborators revealed that humans did not evolve from ancient chimpanzees, as has long been believed. Humans instead evolved along a separate lineage from the last common ancestor shared by early hominids and extinct apes.

Like modern humans, Ardi could walk upright and didn't use her arms for walking, as chimps do. However, she retained a primitive big toe that could have been used in ape-like tree grasping.

Years of field work uncovered Ardi's skull, teeth, arms, hands, pelvis, legs and feet (composite photograph pictured). The paper is accompanied by 10 companion articles.

GENOMICS
Digging out a diabetes gene

Science **324,** 387–389 (2009)

Genome-wide association studies (GWAS) identify genomic regions associated with complex disease, but it can be difficult to pinpoint the exact gene or gene variant involved in the condition. Sergey Nejentsev at the University of Cambridge, UK, and his co-workers show how resequencing certain sections of the genome can bring researchers closer to an answer.

The group resequenced portions of 10 genes — which have been linked to diabetes or related syndromes — in 960 people, half of whom had type 1 diabetes. They then tested these genomic regions' association with diabetes in more than 30,000 people

with and without the disease, and found four rare variants of the gene *IFIH1* that seem to protect against diabetes. This gene is involved in the immune response to RNA viruses. The group concludes that resequencing can help to sort out causative from non-causative genes identified by GWAS.

IMMUNOLOGY
Breaking into the brain

Nature Immunol. **10,** 514–523 (2009)

Immune cells can normally penetrate the barrier between the bloodstream and the brain only when the cells along this barrier are inflamed. So how do immune cells enter the brain to initiate the sort of inflammation associated with autoimmune disorders such as multiple sclerosis?

Federica Sallusto of the Institute for Research in Biomedicine, Bellinzona, Switzerland, and her colleagues showed that the cells enter through the choroid plexus, the brain region where cerebrospinal fluid is made. They used a mouse model of multiple sclerosis to show that a subset of T cells gains entry when their CCR6 surface receptors bind to a protein produced by cells of the choroid plexus. Once inside, the immune cells initiate inflammation.

The team also found that this gateway protein is present in the human choroid plexus and that a higher number of T cells from the spinal fluid of patients with multiple sclerosis express the CCR6 receptor.

BIOCHEMISTRY
Changing bases

Science **324,** 930–935; 929–930 (2009)

Modified versions of the four DNA bases — adenine, thymine, cytosine and guanine — can serve special purposes. Parasitic protozoa called trypanosomes have a modified version of thymine called J that has not been documented in other organisms.

In a study published in May, Anjana Rao of Harvard Medical School in Boston and her colleagues searched for enzymes similar to those responsible for making base J that might generate a similar base in mammals. They found TET1, which converts a methylated cytosine to hydroxymethylcytosine. This altered version accounts for 4–6% of all cytosines in the DNA of mouse embryonic stem cells. The researchers think that this enzyme may regulate gene activity through this modification.
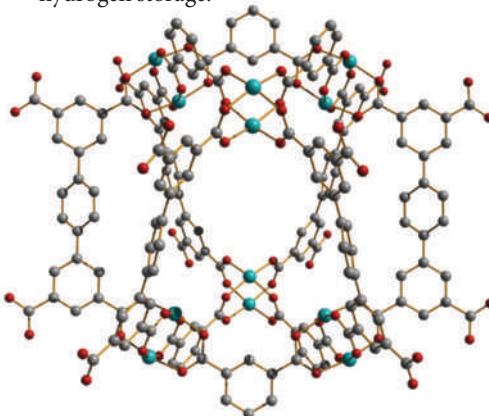
Meanwhile, independently, Skirmantas Kriaucionis and Nathaniel Heintz at the Rockefeller University in New York identified the modified base in the mouse brain.

MATERIALS CHEMISTRY
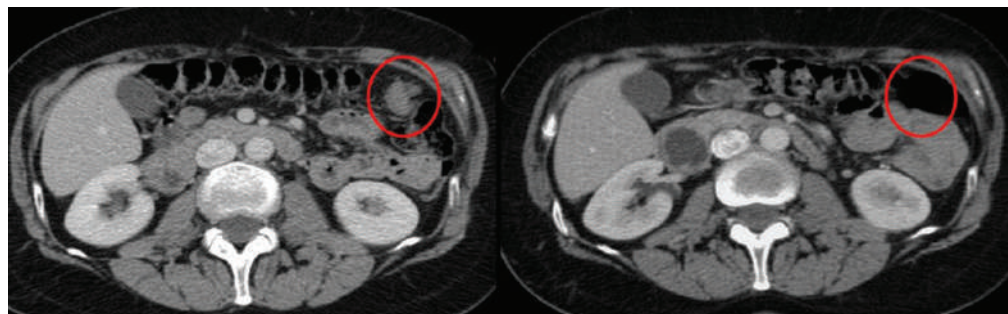# Marvellous metal–organics

J. Am. Chem. Soc. **131,** 2159–2171; 3625–3627 (2009)

Metal–organic frameworks (MOFs) are porous materials with large surface areas that are made up of metals linked through organic groups. They are the focus of research for a range of applications, including hydrogen storage.

Martin Schröder at the University of Nottingham, UK, and his colleagues investigated which parts of a series of copper-based MOFs (one of which is pictured, above) were responsible for increasing the hydrogen-trapping ability of the materials. They showed that the shape of the linking groups and of the other chemical groups around the copper atoms were important. Bare copper sites held on to hydrogen the strongest. Pore size also affected how much hydrogen could be stored. One of the team's materials was able to store 7.78% by weight of hydrogen.

Another study from Anthony Cheetham at the University of Cambridge, UK, and his colleagues, showed that it is possible to make a multiferroic MOF — a material with more than one unique magnetic property. Multiferroic materials are hard to come by because they require the material to take on different molecular arrangements. But the team created a multiferroic MOF from manganese and organic groups, which might be used as magnetic-field sensors and in other devices.

CLIMATE CHANGE
What's up with sea levels?

Glob. Planet. Change **65,** 83–88 (2009)

Global warming causes sea levels to rise in two ways: by melting land ice and by thermally expanding ocean water. In the past, the two factors have proved difficult to accurately tease apart, but Anny Cazenave of the Laboratory for Studies in Space Geophysics and Oceanography (LEGOS) in Toulouse, France, and her colleagues succeeded in measuring the two processes separately. They used gravity data from the GRACE satellites and temperature records collected by the Argo network of buoys.

They calculate that the thermal expansion of sea water has slowed recently and contributed only 0.3 millimetres per year to sea-level rise between 2003 and 2008. Meanwhile, melting land ice caused ocean levels to increase by about 2 millimetres per year.

CANCER DRUG DEVELOPMENT
Targeted tumour take-out

N. Engl. J. Med. **361,** 123–134 (2009)

Tumour cells with a mutation that affects their ability to repair severed DNA can be killed with a drug that knocks out a second repair mechanism. That's the finding from a phase I clinical trial that enrolled patients with cancer-associated mutations in the *BRCA* DNA-repair genes.

Johann de Bono of the Institute of Cancer Research in Sutton, UK, and his co-workers gave patients olaparib, which inhibits an enzyme involved in a second DNA-repair pathway. Out of 19 patients with a *BRCA* mutation, 12 showed signs of tumour shrinkage or stabilization. (Pictured below: left, a computed tomography scan of the abdomen of one patient with ovarian cancer (circled); and right, showing complete tumour regression after four months of treatment.)

The study shows promise for using cancer molecular biology to devise personalized therapies that exploits a tumour's genetic or molecular defects.

2009: THE YEAR IN WHICH …

Words by Lizzie Buchen

H1N1 swept the planet

L. SENG SIN/AP PHOTO

The first influenza pandemic in 40 years propelled the globe onto a roller coaster of panic and complacency. In March, a new H1N1 virus — a mongrel containing genes from swine, bird and human flu viruses — emerged in North America and spread rapidly, sparking fears of a severe pandemic. The new virus was particularly dangerous for younger adults and those with underlying diseases, but most patients had mild symptoms. The low severity cut the world some slack, as a vaccine took months to produce and some manufacturers fell behind schedule; the United States, Australia and Europe didn't start vaccination programmes until October, and poorer nations months after that, if at all. As of mid-December the flu was continuing to intensify across central and eastern Europe and parts of Asia, but its second wave had peaked in North America and parts of Europe. More than 10,580 people have died.

The LHC broke a world record

The high-energy physics crown has passed from the United States to Europe. On 30 November the Large Hadron Collider (LHC) near Geneva, Switzerland, became the highest-energy accelerator in the world, breaking the record long held by the Tevatron at Fermilab in Batavia, Illinois. Europe's proton-pummelling behemoth had spent most of the year in recovery after an electrical failure during its first run in September 2008 caused massive damage. By December 2009, head-on collisions at the LHC had reached 2.36 teraelectronvolts; physicists plan to begin science at this energy level in 2010, with the hope of finding evidence of the long-sought Higgs boson and dark matter.

Climate e-mails were hacked

In what climate-change sceptics are calling the scandal of the decade — and many climate scientists are calling a meaningless nuisance — more than 1,000 e-mails between top researchers were hacked from the Climatic Research Unit (CRU) at the University of East Anglia, UK, just weeks before the Copenhagen climate summit began. Some of the e-mails revealed frustration with data and a cavalier attitude towards sceptics, but they did not discredit the solid body of evidence showing that the world is getting warmer, probably at the hands of humans. The e-mails did, however, embolden sceptics, who interpreted them as evidence of a global conspiracy. CRU director Phil Jones, who composed most of the more controversial e-mails, has stepped aside while an independent panel investigates.

The Moon was found to be damp

A decades-long debate has been resolved: water ice can accumulate in frigid craters on the Moon. On 9 October the rocket booster of NASA's Lunar Crater Observation and Sensing Satellite (LCROSS) slammed into a lunar crater that receives no sunlight, kicking up a plume of dust that was disappointing to spectators but exciting for scientists. Just before LCROSS crashed, its instruments detected water in the dust, suggesting that vapour had frozen into the crater floor. Other spectra hinted at other molecules, such as carbon dioxide, mercury and methane. Researchers hope to explore the ice for clues about the Solar System's history.

Science weathered the recession

The global economic downturn forced tough choices on research funding. Crashing endowments struck top institutions such as Harvard University in Cambridge, Massachusetts, which halted construction on a new science complex, and California's formidable budget deficit drove its state university system to force faculty members to take unpaid leave. But some governments, seeing research as a driver of the economy, made significant efforts to bolster basic science. The US Congress awarded $21 billion in stimulus funding to research. German chancellor Angela Merkel signed off on €18 billion (US$26 billion) over the

They said it …

> **"It's sort of sad isn't it that you criticize government policy and you get sacked?"**
>
> Psychopharmacologist David Nutt after being dismissed from the UK Advisory Council on the Misuse of Drugs. He had advised that alcohol is more harmful than both ecstasy and cannabis.
> *Source: Nature News*

> **"I think the population is losing half of the human brain**

power by not encouraging women to go into the sciences."
>
> Chemistry Nobel prizewinner Ada Yonath, in what turned out to be a record year for women winning the prize: five took the award.
> *Source: AFP*

> **"I have two passions. Space exploration and hip hop."**
>
> Buzz Aldrin, who made the hip hop video

Rocket Experience in the year of the 40th anniversary of the Apollo Moon landings.
> *Source: http://buzzaldrin.com*

> **"In England there has been almost a fashion recently for suing scientists for libel."**
>
> Journalist Simon Singh, who was sued for libel by the British Chiropractic Association, has become the focus of a campaign to keep UK libel law from stifling scientific debate.
> *Source: Australian Broadcasting Corporation*

> **"A high level of serious hysteria."**
>
> Physicist Gordon Kane describes a workshop on a mysterious dark-matter signal (see page 967). The signal was "inconclusive, sadly".
> *Source: New York Times*

> **"The bird escaped unharmed but lost its bread."**
>
> CERN statement after the Large Hadron Collider had an electrical short circuit

next decade to universities and research organizations. French president Nicolas Sarkozy promised to spend €19 billion on research and higher education. And Japan allocated ¥1.6 trillion ($18 billion) for low-carbon technologies. But rich countries looked out for their own: much of African science is hurting after foreign investment and donor funding were slashed.

Japanese science got budget whiplash

Japan's first new government in five decades jolted scientists to attention. In November, a cabinet-level working group chaired by Prime Minister Yukio Hatoyama recommended slashing funding for many prominent science projects, including the SPring-8 synchrotron in Harima, the country's deep-sea-drilling programme and a supercomputer at RIKEN, which would become the world's fastest. Leading Japanese researchers and Nobel laureates rallied the scientific community and appealed to Hatoyama, and in December the country's highest science-policy-making body proposed continued support for the projects. Final budget decisions will be announced by the end of the year.

Climate negotiations faltered

The warmest decade on record concluded with the United Nations climate summit in Copenhagen, where international climate negotiators failed to craft a treaty to succeed the 1997 Kyoto Protocol (see page 966). Deep fissures remained between developing nations, which will probably be affected most by global warming, and developed nations, which have historically emitted the lion's share of greenhouse gases. However, both sides offered some concessions. Rich nations pledged to speed up the delivery of clean-energy technology to the developing world, and developing nations such as Mexico, Indonesia, South Korea and South Africa promised major emissions cuts.

Plagiarism scandal hit Iranian ministers

A series of *Nature* investigations uncovered plagiarism in papers co-authored by high-level officials in Iran. Science minister Kamran Daneshjou, who oversaw the disputed June presidential election, and transport minister Hamid Behbahani, among others, were co-authors on multiple plagiarized papers, most of which have now been retracted by the journals involved.

NASA/MSFC

US human spaceflight was in limbo

NASA may have to lower its ambitions. A commission headed by former aerospace executive Norman Augustine concluded that the US human-space-flight programme has nowhere near sufficient resources to meet the goals laid out by former US President George W. Bush, including returning to the Moon by 2020 using the next-generation Ares rocket (pictured above). Under the agency's current budget, the panel said, astronauts won't even make it beyond low Earth orbit. The report suggested tabling the Moon mission, increasing the role of the private sector and extending the life of the International Space Station, which NASA had planned to abandon in 2015. President Barack Obama met with NASA administrator Charles Bolden on 16 December to discuss the agency's goals, and is expected to announce a decision soon.

MCNAMEE/GETTY

Obama boosted US science

"We will restore science to its rightful place," said US President Barack Obama in his inaugural address in January. On 9 March, Obama (left) signed a memo supporting scientific integrity in federal decision-making and an executive order lifting the prior administration's limits on human embryonic stem-cell research. The latter move greatly expanded the number of cell lines eligible for federal funding for research, and by mid-December, 40 such lines had been approved. Obama has also appointed top scientists to key positions, including physicist Steven Chu as Secretary of Energy (see page 978), physicist and climate expert John Holdren as science adviser and marine ecologist Jane Lubchenco as administrator of the National Oceanic and Atmospheric Administration.

---

when a bird dropped a bit of baguette.
Source: CERN

"I don't want to use a word like 'breakthrough', but I don't think there's any doubt that this is a very important result."
Director of the US National Institute of Allergy and Infectious Diseases Anthony Fauci on a clinical trial that combined two failed HIV vaccine candidates to produce a moderate result.
Source: New York Times

"I'd like to go back to studying cancer epidemiology and aetiology. They're my thing."
Health researcher and former prostitute Brooke Magnanti discusses her future after 'outing' herself as author of the famous sex blog Belle de Jour.
Source: Times Online

"We didn't pay 37 million zlotys ($11 million) for the largest elephant house in

Europe to have a gay elephant live there."
Michal Grzes, a councillor in Poznań, Poland, protests that 'Ninio the gay elephant' will not help the zoo's breeding programme.
Source: Reuters

"Darwin was wrong."
This cover line in *New Scientist* magazine, in an issue marking the 200th anniversary of Darwin's birth, caused much ire.

"I'm not going to answer that question. I am a Christian, and I don't think anybody asking a question about my religion is appropriate."
Gary Goodyear, Canada's minister of state for science and technology, declining to answer the question 'do you believe in evolution?'. His rationale triggered much criticism.
Source: Globe and Mail

NEWS BRIEFING

● POLICY

Physics cutbacks: The funding body that distributes the bulk of UK research money for large-scale physics and astronomy announced drastic cuts on 16 December, in a bid to solve its £40-million (US$64-million) budget gap (see *Nature* **462,** 396; 2009). Researchers responded with disappointment and anger after the Science and Technologies Facilities Council announced a 10% across-the-board cut to grants and a gradual 25% reduction in the number of students and researchers — at present about 280 — that it directly supports with fellowships. A five-year plan also calls for a "managed withdrawal" from more than two dozen individual projects, and £71 million in cuts to ongoing projects. See go.nature.com/1mn2zW for more.

Joint Mars plan: Member states of the European Space Agency (ESA) have formally agreed to team up with NASA to explore Mars in the latter half of the next decade. At an agency council meeting on 16–17 December in Paris, France, ESA's council gave the go-ahead to contribute €850 million (US$1.2 billion) to two missions in the joint-agency 'ExoMars' programme (see *Nature* **460,** 675; 2009). The first is an orbiter to detect trace gases, allied with a small lander, to be launched in 2016, the second consists of two larger planetary rovers to be launched in 2018.

Neutron rivals team up: Hungary has become the fourteenth country to join the Scandinavian-led consortium to build the European Spallation Source (ESS) in Lund, Sweden. The €1.5-billion (US$2.2-billion) ESS will generate high-power neutron beams by firing protons at a heavy-metal target. Hungary had promoted its own rival site for the neutron facility, but earlier this year several key nations threw their support behind the Lund site (see *Nature* **459,** 626;

NOAA/NSF

OCEAN VOLCANO CAUGHT ON FILM

Footage of a volcanic eruption nearly 1,200 metres under the sea was unveiled on 17 December at a meeting of the American Geophysical Union (AGU) in San Francisco, California. The high-definition film of the West Mata volcano in the western Pacific is one of the first times that lava has been caught flowing on the deep sea floor. Jason, the remotely operated vehicle that took the pictures, "actually reached into the molten lava and pulled out a gob" of freshly erupted rocks, says lead researcher Joseph Resing of the Pacific Marine Environmental Laboratory in Seattle, Washington. See go.nature.com/sc837n for more, and go.nature.com/Q2ZovD for all of *Nature*'s reports from the AGU meeting.

2009). Spain, another rival bid, joined the Swedish project in June. Colin Carlile, director of the ESS project, says the next stage — a three-year design review — will begin in 2010.

Innovation clusters named: The European Institute of Innovation & Technology last week named the clusters of European centres that will host its first three 'knowledge innovation communities' (KICs). The scheme is meant to foster greater research collaboration between academia and industry across Europe. Focusing on climate change, sustainable energy and information technology, each of the three KICs will get €3 million (US$4.3 million) in start-up funding and will try to win more support from national and European funding programmes, and private sources.

SOUND BITES

"Meaningful and unprecedented."

US President Barack Obama puts a positive face on the Copenhagen accord. See page 966 for more.

● BUSINESS

Solar going public: Solyndra, a photovoltaics manufacturer based in Fremont, California, filed with the US Securities and Exchange Commission on 18 December for an initial public offering. The firm wants to raise up to US$300 million. Founded in 2005, Solyndra makes thin-film cylindrical modules for use in solar arrays on commercial roof-tops. In March 2009, it was granted a $535-million loan guarantee from the US government to expand manufacturing capacity.

Carbon permits fall: Reacting to the climate accord announced in Copenhagen (see page 966), prices for permits to emit a tonne of carbon dioxide (for delivery in 2010) closed at a 6-month low of €12.45 (US$17.78) on the European Climate Exchange

on 21 December. This is well down on prices that were solidly above €14 in the week before the conference. Traders think that the uncertainty associated with the accord will see transaction volumes and prices continue to slide.

Vaccine hire: Julie Gerberding, who was director of the US Centers for Disease Control and Prevention in Atlanta, Georgia, during 2002–09, has been hired to head the vaccines unit of pharmaceutical giant Merck & Co. She stepped down from the health agency when US President Barack Obama took office on 20 January 2009.

NASA/C. THOMAS

● RESEARCH

Stem-cell gap: The human embryonic stem-cell lines in wide use among researchers are strikingly lacking in population diversity, according to a report published last week (J. T. Mosher *et al. N. Engl. J. Med.* doi:10.1056/NEJMc0910371; 2009). University of Michigan scientists analysed the genetic ancestry of 47 commonly used lines — including 22 approved for US government funding earlier this month. They found that 31 were of exclusively northern and western European origin, 2 were Asian, and the rest had elements of southern European or Middle Eastern ancestry, or both. None was of African origin.

Geologist acquitted: A Swiss court on 21 December cleared geologist Markus Häring of any wrongdoing relating to small earthquakes caused by an

NUMBER CRUNCH
20%

The contribution of aircraft vapour trails to warming in the Arctic since temperature records began.

Source: Mark Jacobson, Stanford University, Palo Alto, California. American Geophysical Union meeting in San Francisco, California. See go.nature.com/Txwdhf for more.

innovative geothermal drilling project in Basel, Switzerland, which he had led with the Geopower Basel consortium. He had not played down any risks and acted in accordance with the state of scientific knowledge at the time, the court said. The project was shut down on 10 December following a government study into the economic damage it had caused (see also *Nature* **462,** 848–849; 2009).

Airborne telescope on test: A decade behind schedule, the Stratospheric Observatory for

Infrared Astronomy (SOFIA) has finally been unshuttered. On 18 December, a door in the side of the Boeing 747 that houses the telescope was opened during a flight (pictured) — exposing the 2.5-metre telescope to the atmosphere for two minutes. An audit earlier this year (see *Nature* **458,** 693; 2009) found that costs of the project had more than tripled from the original estimate of US$265 million. NASA, which funds SOFIA together with the German Aerospace Center, says scientists' first opportunity to use the telescope could come in spring 2010.

Autism alert: The prevalence of autism in the United States has grown dramatically, to 1 in 110 children, according to an 18 December report by the US Centers for Disease Control and Prevention in Atlanta, Georgia. The numbers, published in *Morbidity and Mortality Weekly Report*, reflect a 57% surge in cases of autism spectrum disorder (ASD) in 8-year-olds between 2002, the last survey year, and 2006. "Although improved ascertainment accounts for some of the prevalence increases documented … a true increase in the risk for children to develop ASD symptoms cannot be ruled out," the authors wrote.

THE WEEK AHEAD

1 JANUARY 2010
Spain assumes a six-month presidency of the European Union.

3–7 JANUARY 2010
The American Astronomical Society meets in Washington DC.
▶ http://aas.org/meetings/aas215

3–7 JANUARY 2010
The Society for Integrative and Comparative Biology meets in Seattle, Washington.
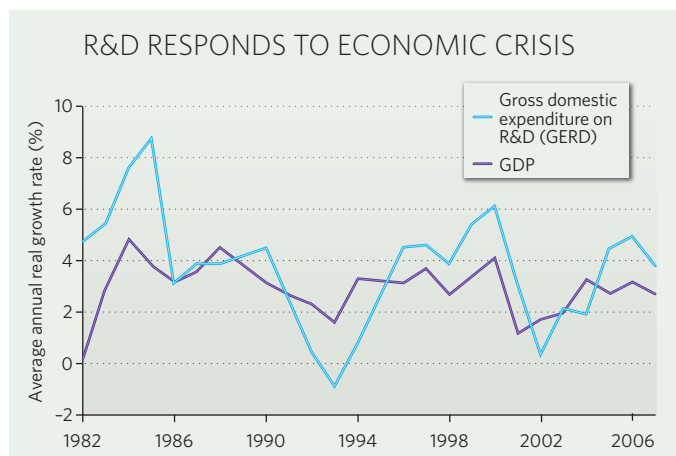▶ www.sicb.org/meetings/2010

● FUNDING

Neglected diseases funding: Spending on research and development into neglected diseases "ground to a standstill" in 2008, according to the second edition of the report *Global Funding of Innovation for Neglected Diseases*, published by the George Institute for International Health in Sydney, Australia. Donors, governments and industry combined spent US$2.96 billion on new products for these diseases. Although that was around 4% up on 2007, most of this rise was due to funds from the Bill & Melinda Gates Foundation, based in Seattle, Washington, which increased its spending by 36.5% to $617 million. This masked an overall trend of freezes or slight falls in support from rich countries and most donors.

BUSINESS WATCH

Preliminary data from US stock markets suggest that companies have already reduced their investments in research and development (R&D) following the economic crisis. More than 2,000 companies that filed quarterly reports with the US Securities and Exchange Commission reported an average reduction of 6.6% in their planned R&D spending in the first quarter of 2009. The data are contained in the Organisation for Economic Co-operation and Development's *OECD Science, Technology and Industry Scoreboard 2009*, officially released on 14 December. Vincenzo Spiezia, a senior economist at the OECD, says actual spending may drop even lower. Historically, R&D spending lags behind gross domestic product (GDP; see chart). Spiezia thinks that the lowest point for R&D has not yet been reached, and notes that only a small proportion of the economic-stimulus packages went directly to bolster research. The OECD also noted a huge drop in venture-capital investment in the United States — another indicator of dulled R&D prospects.

R&D RESPONDS TO ECONOMIC CRISIS

Average annual real growth rate (%)

— Gross domestic expenditure on R&D (GERD)
— GDP

SOURCE: OECD

NEWS

World looks ahead post-Copenhagen

A weak international climate agreement leaves room for science to shape the next round of negotiations.

COPENHAGEN

Two lines of evidence nearly brought down the last-minute climate agreement brokered last week in Copenhagen by US President Barack Obama: studies indicating that the impacts of global warming could be more severe than previously thought, and that rich countries could do more to counter the problem without breaking the bank.

Now, negotiators are seeing whether they can strengthen a deal nearly universally acknowledged as weak — or whether even the mounting scientific evidence on the most dire effects of climate change will be enough to forge a more meaningful deal.

The Copenhagen accord, drawn up as a multilateral political agreement between the United States, China, India, Brazil and South Africa, provides a non-binding framework for capturing national commitments that are already on record. Many environmentalists and a number of developing countries criticized it immediately, saying that current climate commitments would not meet the common goal of limiting the increase in global average temperatures to 1.5–2 °C. Indeed, an analysis by the US modelling consortium Climate Interactive estimated that the commitments would result in a temperature rise of 3.9 °C by the year 2100.

For its part, the European Union elected in Copenhagen not to increase to 30% its conditional offer of a 20% cut in 1990 emissions levels by 2020. The overall goal remains to hammer out a treaty to replace the 1997 Kyoto Protocol on Climate Change while incorporating climate commitments from developing countries and the United States, which has not ratified the Kyoto agreement.

The Copenhagen accord was blasted by environmental activists and leaders of developing nations.

A. KISBENEDEK/AFP/GETTY

Negotiations will continue over the coming year, and many observers hope for a stronger deal when the parties to the United Nations Framework Convention on Climate Change meet in Mexico in late November 2010.

Obama's announcement on the last day of the Copenhagen talks fractured the G77 group of developing countries. Lumumba Stanislaus Di-Aping, Sudan's fiery chief negotiator and chairman of the G77, claimed that the deal would do nothing less than "destroy" Africa. He was joined by countries including Tuvalu, Bolivia and Venezuela during an all-night debate that included the occasional tirade against capitalism but also focused on how poor nations would be affected disproportionately by climate change (see 'Growing agricultural benefits for climate'), and how rich countries should be paying more.

For instance, scientists at the Copenhagen meeting had enlisted the star power of Al Gore in releasing a pair of reports on the accelerated

Growing agricultural benefits for climate

Building on the momentum of using carbon markets to protect forests, scientists are looking at ways to integrate agriculture into climate-change discussions. Better seeds and sustainability practices could make farming more resilient to climate change while increasing yields, cutting emissions and reducing pressure to clear land.

Agricultural practices are responsible for about 15% of greenhouse-gas emissions but are the main driver of emissions from deforestation. In Copenhagen, negotiators made progress on a programme known as REDD (Reducing Emissions from Deforestation and Forest Degradation). But "if you don't improve agriculture, you'll never get REDD right", says Sandra Brown, a forest-biomass expert with the non-profit organization Winrock International in Arlington, Virginia.

Pressure on agriculture will increase in a warming world with more people. A recent study by the United Nations World Food Programme found that shifting precipitation patterns caused by global warming could increase the number of malnourished children by 21% in 2050; some 10 million of those would be in sub-Saharan Africa. Reporting similar numbers in Copenhagen, scientists with the Consultative Group on International Agricultural Research in Washington DC said that the problem can be addressed by producing better seeds, improving agricultural practices, expanding irrigation and other infrastructure and helping farmers switch to drought-tolerant crops .

"With appropriate expenditures, we can actually address the impacts of climate change," says Thomas Roswell, an ecologist at the Swedish University of Agricultural Sciences in Uppsala. **J.T.**

loss of glaciers and ice sheets in Greenland and Antarctica; melting ice sheets in particular, they concluded, could increase sea level by at least 1 metre by 2100.

Easy targets

Still, an economic analysis by the International Institute for Applied Systems Analysis (IIASA) in Laxenberg, Austria, found that the pre-Copenhagen commitments by industrialized countries would reduce 2020 emissions by 11–22%. Taking into account cost savings due to efficiency and other factors, the cost of achieving that goal would be just 0.15% of gross domestic product, the analysis revealed. Markus Amann, who heads the IIASA's greenhouse-gas initiative, fears that such easy targets will translate into low carbon prices and end up delaying the innovation needed to make deep cuts in decades to come.

Yet most developing countries ultimately defended the Copenhagen accord — with reservations — because without it there would have been nothing to show for the largest environmental conference in history. "This is not what we have been seeking, but it is a beginning," said Mohamed Nasheed, president of the Maldives, which is a member of the Alliance of Small Island States that has been pushing for a commitment to limit warming to 1.5 °C. "I beg all nations to please back this document and do not let these talks collapse."

"The most important thing to get done at this moment is to get moving," adds John Holdren, Obama's chief science adviser. "Let's not argue for the next five years what the perfect goal is. Let's get going in the right direction."

For the first time under the new framework, both developing and developed countries — including the United States, which is gearing up for a legislative battle over climate in the Senate this spring — will be bound under a single agreement. Rich countries would be required to meet their promised reductions; developing countries would be required to audit and report the activities they undertake to reduce emissions, as well as open their books to international verification for any projects funded with international aid.

In exchange, the agreement offers financing, valued at US$30 billion by 2012, to help developing countries prepare for a warmer world and to push forward with sustainable develop-

ment goals. Developed countries committed to spending upwards of $100 billion annually by 2020, although the agreement offers no details as to where that money will come from. The World Bank has estimated that developing countries will need about that amount, but developing countries and many scientists say that number could be far too low.

Martin Parry of Imperial College London, a former co-chair of the Intergovernmental Panel on Climate Change (IPCC) working group on impacts, adaptation and vulnerability, has developed a way to portray 'unavoided impacts' in which mitigation fails and there isn't enough money to help people cope with the consequences. Assuming that all countries fulfil their pledges, the agreement in Copenhagen still leaves a gap of 1.5 °C, he says. In other words, funding for adaptation could cover impacts associated with about 1.5 °C of warming, but temperatures are likely to rise by at least 3 °C.

"Even the toughest and most robust measures will not achieve 2 °C," Parry says. "We should hope for 2 °C but realistically expect 3–3.5 °C, and then plan for 4 °C."

The Copenhagen accord includes a goal of limiting temperature rise to 2 °C, but one target that was notably dropped was that for reducing global greenhouse-gas emissions by 2050. Small-island states successfully pushed for language requiring consideration of a target of 1.5 °C during the first review of implementation and commitments in 2015, one year after the IPCC is scheduled to finalize its fifth assessment report. Climate modelling for the next IPCC assessment is under way, and the panel will accept nominations and then appoint lead authors for the various chapters this spring.

Sitting in his office overlooking the main hall of the Bella Center in Copenhagen last week, IPCC chair Rajendra Pachauri said that he was encouraged by the fact that so many world leaders chose to attend, even if a commitment to act falls short.

"There's a certain inertia that will resist it. There are mindsets that will resist it. There are vested interests that will resist it — let's not minimize their effectiveness," Pachauri says. "This is something that politically one had to anticipate." ∎

Jeff Tollefson
See Editorial, page 957.
For more, see www.nature.com/roadtocopenhagen.

I. KALNINS/REUTERS

Delegates worked through the night to reach an agreement.

'Dark-matter' events spotted

Two sightings in Minnesota have set physicists buzzing about whether the first direct detection of dark matter has been made. If confirmed, it would mark the end of a decades-long search for the mysterious particles thought to make up as much as 85% of matter in the Universe.

But most agree that the signals are not statistically significant enough to be attributed to dark matter rather than to conventional particles.

The two events were caught in 2007 in super-cooled crystals of germanium and silicon in the underground Cryogenic Dark Matter Search II (CDMSII) experiment in the Soudan Mine in Minnesota. Last week, CDMSII scientists announced that they have seen candidates for the dark-matter particles known as weakly interacting massive particles (WIMPs), each with a mass of 30–60 gigaelectronvolts — roughly 30–60 times that of a proton.

But from the analysis, team scientists think that there is a 25% chance that both events might be false-positives caused by background radiation. Those odds are not good enough to claim a definitive detection of WIMPs, says Timothy Sumner, a physicist at Imperial College London. "Statistically, it's not compelling," he says.

"The best we could call it is a hint," adds John Ellis, a theoretical physicist at CERN, Europe's high-energy physics lab near Geneva, Switzerland. "An interesting hint."

The possible detection is the latest in a series of potential dark-matter sightings. In August 2008, an Italian-led satellite-based experiment known as PAMELA reported an excess of antielectrons (positrons) that could have stemmed from the annihilation of dark-matter particles. And in October 2009, NASA's Fermi Gamma-ray Space Telescope saw a haze of high-energy light in the centre of our Galaxy that could be a dark-matter signature.

The CDMSII result will now spur physicists at the Large Hadron Collider (LHC) at CERN to try to generate WIMPs in their collisions. "The LHC would see this very easily and relatively quickly," says Ellis — and could potentially produce a detectable WIMP signal by the end of next year. ∎

Geoff Brumfiel
For a longer version of this story, see go.nature.com/34rnga.

Tsunami watch

Five years after the Indian Ocean disaster, the technology is in place, but local preparedness is less advanced.

On the morning of 29 September, Mase Akapo knew exactly what to do. At 6:48 a.m., on the island of Tutuila in American Samoa, he had felt the ground shake harder than he ever had before. Ten minutes later Akapo, a meteorologist, was in his office, issuing an alert for the tsunami he knew was probably on its way.

Tsunami warnings are a way of life for coastal communities, but five years ago they took on a new layer of meaning. On 26 December 2004, a magnitude-9.0 quake off the Indonesian island of Sumatra generated a series of tsunamis that drowned some 230,000 people around the Indian Ocean. Most of them had no idea it was coming.

Memory of the disaster remains relatively fresh in many coastal communities, even in the South Pacific, where American Samoa and other nations have used it to brush up on their preparedness plans. "What we learned from that tsunami really helped us," says Akapo, whose 2009 warnings helped entire villages to reach safety before a massive wave struck Tutuila.

Experts say that much work remains, however. In all the world's major oceans — although not the Mediterranean Sea, where some still see a risk — sophisticated tsunami-sensing instruments are now on alert, from the extensive Pacific network first set up in 1965 to the brand-new system deployed across the Indian Ocean in the wake of the 2004 disaster (see graphic). Such systems rely on a network of seismic stations to detect the earthquake, and deep-ocean and coastal gauges to detect resulting changes in sea level. But the best instrumentation in the world cannot guarantee that crucial communication takes place where it is needed: at the waterfront, before the wave strikes, and in terms that local communities can understand and heed.

"In the Indian Ocean region there is a tremendous amount of work yet to be done," says Costas Synolakis, director of the University of Southern California's Tsunami Research Center in Los Angeles. "It is urgent work, because when it comes to tsunamis, bad information kills."

Before 2004, most people associated tsunamis with the Pacific Ocean, where the waves have repeatedly struck Japan and Hawaii. Few worried about such hazards in the Indian Ocean. Only 4% of all known tsunamis in the twentieth century occurred there — and none had struck in living memory in countries such as Thailand, India and Sri Lanka. Palaeorecords of ancient tsunamis suggest that the 2004 event was the biggest in that region in more than 600 years[1,2].

At the time, the Indian Ocean had no tsunami warning system. By 2005, the Japan Meteorological Agency in Tokyo and the Pacific Tsunami Warning Center in Honolulu, Hawaii, had started issuing warnings as well as they could for the Indian Ocean with the instruments they had set up for the Pacific. And by the end of March 2010, the new Indian Ocean Tsunami Warning and Mitigation System will be fully operational. It includes 120 seismic stations, more than 60 coastal tide-gauge stations, 24 early-detection buoys and 20 deep-ocean 'tsunameters'.

The effort, which cost more than US$100 million, is being paid for by various countries, with Germany, Indonesia and India assuming the largest shares. The 28 countries that are connected to the system are each responsible for collecting and feeding in their data to regional tsunami watch centres in Australia, India and Indonesia. If the data suggest a real threat, the centres will immediately send an alert back to national tsunami focal points, which in turn disseminate warnings to local communities and emergency services.

"People in the region are safer than they were in 2004," says Keith Alverson, project-office director with the Global Ocean Observin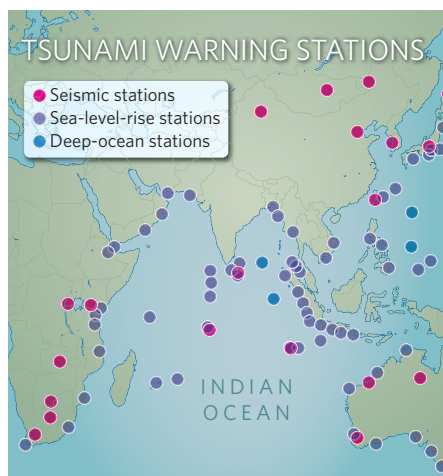g System of the United Nations Educational, Scientific and Cultural Organization (UNESCO) in Paris. "The challenge is to tailor technology to local cultures and make the system sustainable in the long term," he says.

Hidden dangers

A tsunami that struck Java on 17 July 2006, killing more than 600 people, highlighted the limits of any warning system, no matter how technologically advanced. The tsunami arrived about 30 minutes after the earthquake, and survivors have reported that no warnings were issued. Lifeguards also failed to recognize the initial recession of water that typically precedes a tsunami, because it was masked by large, wind-driven waves breaking on the shore[3].

The earthquake itself might have raised the alarm. "For people living close to a fault line, the only effective tsunami warning, alas, is the quake itself," says Harald Spahn, a geologist with the German Technical Cooperation who helps authorities in Sumatra, Java and Bali to improve tsunami warning capacities at the local level. But before the Java tsunami little or no ground-shaking was felt, which is typical for the 'slow' earthquakes that are common in the region.

Community-based public education and evacuation drills are essential to save lives in

> **"When it comes to tsunamis, bad information kills."**

Early warnings saved lives in a 2009 tsunami in Samoa.

UNESCO, 2008

TSUNAMI WARNING STATIONS

- Seismic stations
- Sea-level-rise stations
- Deep-ocean stations

INDIAN OCEAN

ENCELADUS PLUME IS HALF ICE
Geysers not fed by misty water vapour after all.
go.nature.com/Hpah3H

NASA/JPL/SPACE SCI. INST.

NEW ZEALAND DEFENCE FORCE, NZ HERALD/AP PHOTO

any tsunami, even if evacuation begins just minutes before the wave arrives. Regular drills are now conducted in some Javanese communities, but not all Indonesian coastlines are sufficiently prepared, says Spahn.

There have been some success stories in improving the communication of warnings down the critical 'last mile', he adds. For example, loudspeakers on mosques that normally call people for prayer are also an efficient way of broadcasting a tsunami warning, and have been used for this purpose in Sumatra and Java.

Similar public-education efforts paid off in saving lives in Samoa and American Samoa three months ago, says Bruce Jaffe, a coastal-hazards researcher at the US Geological Survey in Santa Cruz, California. "Many people knew what to do," he says; they hurried on foot to higher ground.

The 2004 Indian Ocean tsunami, he says, provided "the wake-up call" to reinvigorate overall preparedness efforts for coastal communities around the world. But as more years pass since nearly a quarter of a million million people died, Jaffe worries that people's awareness will start to drop off again. "We can't let that happen," he says. "We've got to, if anything, step up the efforts." ■

Quirin Schiermeier, with additional reporting by Alexandra Witze

1. Jankaew, K. et al. Nature 455, 1228–1231 (2008).
2. Monecke, K. et al. Nature 455, 1232–1234 (2008).
3. Fritz, H. M. et al. Geophys. Res. Lett. 34, L12602 (2007).

Extortion attempt involved retracted papers

Two papers retracted in the past few months have been linked to an extortion attempt. Both papers originated from the laboratory of Peter Schultz, a prominent chemist at the Scripps Research Institute in La Jolla, California.

Documents seen recently by *Nature* show that in 2007, law-enforcement officials in San Diego considered a former postdoctoral fellow from the Schultz lab as a possible suspect after another received an anonymous e-mail demanding a $4,000 payment and threatening to reveal alleged fraud.

Law-enforcement officials did not pursue the case after the recipient of the e-mail decided not to press charges.

The retracted papers were published in *Science*[1] and the *Journal of the American Chemical Society* (*JACS*)[2]. They claimed to describe the successful incorporation of amino acids linked to sugars at specific positions in proteins made by the bacterium *Escherichia coli*[3].

On 1 March 2007, Zhiwen Zhang, the first author on the *Science* paper and the third author on the *JACS* paper, received an e-mail that read in part: "you have fraud on at least 3 papers and you stole library material- I found proof." The author of the e-mail asked Zhang to send cash to a post-office box in San Diego, and threatened that if Zhang did not comply, e-mails would be sent to Schultz, Scripps president Richard Lerner, and other scientists and administrators at Scripps and at the University of Texas at Austin, where Zhang began working after he left the Schultz lab in 2004.

Zhang says he did not commit fraud. "I did no wrong, no scientific misconduct and no fraud," he says. "I am the victim of an extortion case, and I have suffered dearly."

The e-mail was forwarded to law-enforcement officials in San Diego. In March 2007, a southern California multi-agency task force, the Computer and Technology Crime High-Tech Response Team, obtained warrants to search the records of Internet service providers in connection with the case. On 6 April, an officer with the task force notified Zhang that it considered Eric Tippmann, who overlapped with Zhang while both were

postdoctoral fellows in Schultz's lab in 2004, to be a suspect. Zhang says that after consulting with Schultz and Lerner, he decided not to press criminal charges. Law-enforcement officials have confirmed that the investigation is closed.

Tippmann denies sending the extortion e-mail or contacting Zhang after Zhang left the Schultz lab. "If I was ever briefly mentioned in any investigation, I was never contacted nor interviewed, so this must have been a very short investigation," Tippmann wrote in an e-mail.

In August of this year, Tippmann and his colleagues published a study[4] claiming that the experiments reported in the now-retracted papers could not have worked as described. Tippmann says he first became concerned about the papers in 2006 after he noticed what he alleges are similarities between mass spectra shown in the retracted *JACS* paper[2] and in an earlier *JACS* paper[5] from the Schultz lab. Tippmann, now at Cardiff University, UK, says he also noticed other inconsistencies in some of the lab's papers, and flagged these to Schultz.

Zhang calls the issues raised by Tippmann "irrelevant". "It has all been cleared up," says Zhang, who says that Scripps has looked into the matter and cleared him of fraud and misconduct.

Schultz says that he has reproduced all of the results questioned by Tippmann, including other work that Zhang was involved in, except for the experiments that have now been retracted. He says that he is unsure why the mass spectra in the 2003 and 2004 *JACS* papers contain some similarities. "My guess is in that case a mistake was made," Schultz says. "I don't think fraud was committed."

Zhang says that he stands by his original work. ■

Erika Check Hayden, with additional reporting by Rex Dalton

1. Zhang, Z. et al. Science 303, 371–373 (2004).
2. Xu, R. et al. J. Am. Chem. Soc. 126, 15654–15655 (2004).
3. Check Hayden, E. Nature 462, 707 (2009).
4. Antonczak, A. K., Simova, Z. & Tippmann, E. M. J. Biol. Chem. 284, 28795–28800 (2009).
5. Alfonta, L., Zhang, Z., Uryu, S., Loo, J. A. & Schultz, P. G. J. Am. Chem. Soc. 125, 14662–14663 (2003).

969

Fraud rocks protein community

University finds that researcher falsified data supporting 11 protein structures.

The finding by a university misconduct investigation that a crystallographer "more likely than not" faked almost a dozen protein structures has left the field in shock. The fraud is the largest ever in protein crystallography. The disputed structures had important implications for discovering drugs against dengue virus and for understanding the human immune system.

"It's massive," protein crystallographer Wayne Hendrickson of Columbia University in New York says of the investigation's conclusion. "It's the worst possible thing."

In a report released earlier this month, the University of Alabama at Birmingham concluded that H. M. Krishna Murthy acted alone in fabricating and falsifying results that appeared in ten papers[1–10] published during the past decade. The disputed papers have been cited more than 450 times.

Murthy denies any wrongdoing. Girish Kotwal, a co-author of Murthy's who was suspended by the University of Cape Town, South Africa, in 2006 owing to charges of professional misconduct (see *Nature* doi:10.1038/news060703-13; 2006) and now runs Kotwal Bioconsulting in Louisville, Kentucky, says that Murthy "feels defenceless and unfairly treated by some in the crystallography field and his institution". Kotwal sent *Nature* a statement that he says was e-mailed to him from Murthy indicating that Murthy disagrees with the findings of the committee and stands "by all of the reported results in these papers, as well as the experimental origin of the underlying structures".

But for the investigation, Richard Marchase, the university's vice-president of research, says that Murthy did not retain a lawyer and "was not able to produce any compelling evidence as to how he arrived at the structures".

All of the disputed structures had been deposited in the Protein Data Bank (PBD). So far, only the dengue virus NS3 serine protease has been both removed from the PDB and retracted by *The Journal of Biological Chemistry*, where it was first published in 1999 (ref. 1). The results in that paper sent the hunt for drugs against this protease down a blind alley. Stanley Watowich, a virus expert at the University of Texas Medical Branch in Galveston, says that

The first of the protein structures to be disputed, that for human C3b.

two of Murthy's structures[1,3,4] were among 14 included in a virtual dengue drug-screening project run over the past year. This modelled how candidate molecules would interact with dengue proteins, using IBM's World Community Grid — a public computing network set up to harness unused computer time for projects of benefit to humanity. "Screening against the Murthy structures took about two months," says Watowich, "and it is unfortunate that this time could not have been more productively spent."

Murthy began his postdoctoral training in the art of protein crystallography — growing crystals and diffracting X-rays through them for clues to structure — as a postdoc at Yale University in New Haven, Connecticut, in 1981. There, he worked in the lab of Thomas Steitz, a crystallographer who this year shared the Nobel Prize in Chemistry for his work on the structure of the ribosome. Murthy joined Hendrickson's lab in 1985, and struck the senior scientist as being a "very solid guy" who had some bad luck. "He grew his crystals, went to the synchrotron, and they didn't diffract very well so he didn't have any fantastic accomplishments out of this," Hendrickson says. He adds that he believes that Murthy did some genuine work in his lab.

Murthy arrived at Alabama in July 1998 to take up a position as a research assistant professor at the Center for Biophysical Sciences and Engineering. The first questions about his work arose with the October 2006 publication in *Nature* of the structure of the human C3b complement-system component, part of the cascade of immune-system proteins that destroys invading cells[10]. A number of groups had been pursuing the structure, and the journal published Murthy's paper alongside similar papers from Bert Janssen at Utrecht University in the Netherlands and his co-workers, and a group from Genentech in South San Francisco[11,12].

When the structures were deposited in the PDB, Janssen immediately noticed discrepancies between Murthy's and his own, including large 'gaps' in the lattice that were unusual in such a well resolved and ordered structure.

Janssen and his supervisor, Piet Gros, enlisted two well known crystallographers, Randy Read of the University of Cambridge, UK, and Axel Brünger of Stanford University, California, to examine it. They agreed that Murthy's structure seemed to be fake. The group sent a brief communication to *Nature* in December 2006 questioning the structure[13] and forwarded their concerns to the University of Alabama.
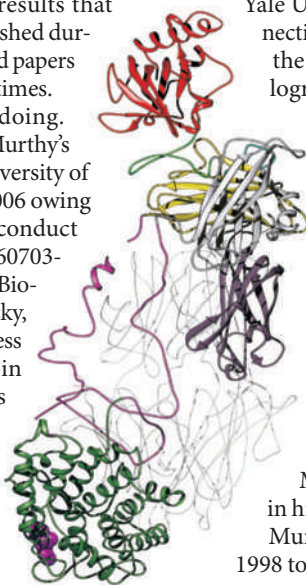
In January 2007, the University of Alabama began a two-year investigation, which reported earlier this month that Murthy had acted alone in fabricating that structure and ten others. How Murthy fabricated data is unclear, but one method he might have used involves grafting the sequences of target proteins onto structures for similar proteins, then using algorithms to back-calculate diffraction intensities, adding realistic errors along the way.

The PDB says it will remove the other ten structures only when editors at the journals in which they were originally published or the authors themselves retract them. Until Murthy's case came along, it had never removed structures from its database for reasons of misconduct.

Shortly after the publication of their *Nature* correspondence, Read and Brünger formed a validation task force at the PDB to provide an automated and confidential means of verifying structures during peer review. "With this validation," Brünger says, "this information will be given to the reviewer and if there are any questions one can go back and request the data." The next disputed protein structure may not take so long to uncover. ■

Brendan Borrell

1. Murthy, H. M. K., Clum, S. & Padmanabhan, R. *J. Biol. Chem.* **274,** 5573–5580 (1999).
2. Urs, U. K., Murali, R. & Murthy, H. M. K. *Acta Crystallogr. D* **55,** 1971–1977 (1999).
3. Murthy, H. M. K., Judge, K., DeLucas, L., Clum, S. & Padmanabhan, R. *Acta Crystallogr. D* **55,** 1370–1372 (1999).
4. Murthy, H. M. K., Judge, K., DeLucas, L. & Padmanabhan, R. *J. Mol. Biol.* **301,** 759–767 (2000).
5. Murthy, K. H. M. *et al. Cell* **104,** 301–311 (2001).
6. Kumar, M. S., Carson, M., Hussain, M. M. & Murthy, H. M. K. *Biochemistry* **41,** 11681–11691 (2002).
7. Ganesh, V. K., Smith, S. A., Kotwal, G. J. & Murthy, K. H. M. *Proc. Natl Acad. Sci. USA* **101,** 8924–8929 (2004).
8. Ganesh, V. K., Muthuvel, S. K., Smith, S. A., Kotwal, G. J. & Murthy, K. H. M. *Biochemistry* **44,** 10757–10765 (2005).
9. Ajees, A. A., Anantharamaiah, G. M., Mishra, V. K., Hussain, M. M. & Murthy, H. M. K. *Proc. Natl Acad. Sci. USA* **103,** 2126–2131 (2006).
10. Ajees, A. A. *et al. Nature* **444,** 221–225 (2006).
11. Janssen, B. J. C., Christodoulidou, A., McCarthy, A., Lambris, J. D. & Gros, P. *Nature* **444,** 213–216 (2006).
12. Wiesmann, C. *et al. Nature* **444,** 217–220 (2006).
13. Janssen, B. J. C., Read, R. J., Brünger, A. T. & Gros, P. *Nature* doi:10.1038/nature06102 (2007).

REF: 10

IMAGES OF THE YEAR

Take a tour through the images that defined 2009, from the murkiest depths of the oceans, where warring worms pepper their enemies with glowing bombs, to the spectacular swirling rainbows thousands of light years away at the heart of the Milky Way.

This year has provided a pictorial panoply, with dust-filled volcanic eruptions captured on film by luck, the tiniest of toads at risk of being lost forever and humankind's outpost in space picked out in detail as it passes in front of the Sun.

Researched and written by Katharine Sanderson.

SPITZER, CHANDRA AND HUBBLE TEAM UP

The Hubble Space Telescope got a new lease of life this year, when astronauts from NASA's space shuttle *Atlantis* fitted the 19-year-old satellite with two new instruments and repaired two old ones. This view of the Milky Way's centre combines near-infrared data from the reinvigorated Hubble (yellow) with infrared signals from the Spitzer Space Telescope (red) and X-ray data from the Chandra X-ray Observatory (blue and violet).

MAGNIFICENT MANDELBROT

Maths shows its beautiful side in the first true three-dimensional representations of the Mandelbrot set. Created by computer programmer Daniel White from Bedford, UK, this is the eerily beautiful 'Mandelbulb'.

SARYCHEV VOLCANO

In June, the International Space Station happened to be in orbit right above the Sarychev volcano on Russia's Kuril Islands, northeast of Japan, as it began to blow. The explosion sent plumes billowing 16 kilometres into the sky, spreading ash across a radius of hundreds of kilometres.

SATURN'S GIANT RING

This artist's impression shows the huge ring discovered around Saturn. Visible only in the infrared spectrum, the ring was spotted by the Spitzer Space Telescope. Thought to be made up of scattered dust particles, the ring has a diameter of at least 25 million kilometres, and might be responsible for darkening one face of Saturn's moon Iapetus.

SMASHING SUCCESS

If 2008 was the Large Hadron Collider's *annus horribilis*, 2009 has seen its redemption. By the end of this year the collider was up and running again after it broke down in 2008, and is now smashing records as well as particles — accelerating protons to the highest energy ever.

LOW SPIRIT

Spirit, one of NASA's two vehicles still exploring Mars after almost six years, has spent much of the past year stuck in some soft soil. The robot is now making tentative escape attempts.

ATOMIC DETAIL

This is a molecule of pentacene, showing five rings of carbon in glorious atomic detail thanks to a specially designed atomic force microscope developed by Gerhard Meyer and his colleagues at IBM Research in Zurich, Switzerland.
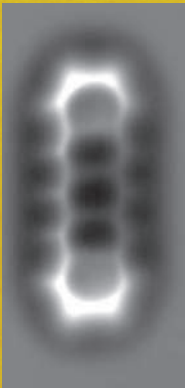
MELTDOWN

This radar image, 150 kilometres across, shows an ice bridge collapsing away from the Wilkins ice shelf in the Antarctic in April. Its disappearance threatens to accelerate the shelf's destruction and allow more ice from the continent's glaciers to flow into the ocean.
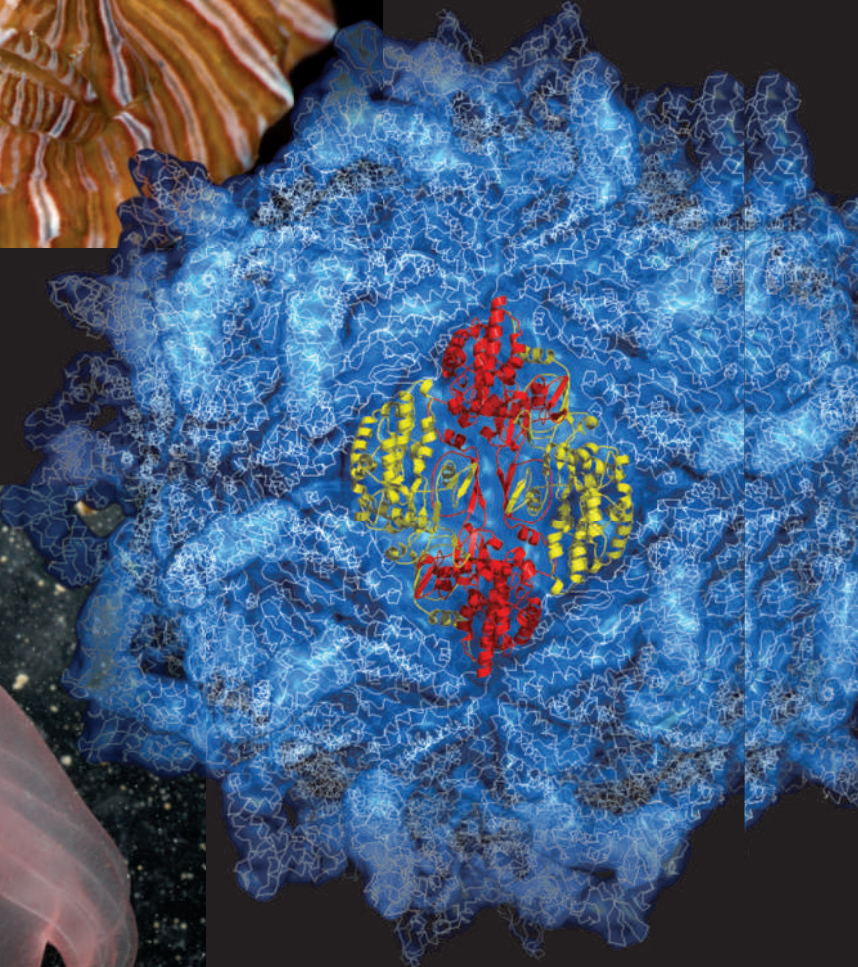
GROOVY

Meet the psychedelic frogfish (*Histiophryne psychedelica*), which was discovered by Theodor Pietsch at the University of Washington, Seattle, and his colleagues, in waters off Indonesia, and gets around with a lolloping hop along the sea floor.

CAPTIVATING CAPSID

It took three years to image the five million atoms in this capsid, a protein coat used by many virus particles to protect their DNA. Junhua Pan, then at Rice University in Houston, Texas, created the picture from hundreds of smaller X-ray-diffraction images of the *Penicillium stoloniferum* virus F, which infects the fungus that makes penicillin.

MYSTERIES OF THE DEEP

This see-through sea cucumber of the genus *Enypniastes* was discovered this year lurking 2,750 metres below sea level in the northern Gulf of Mexico. Part of the Census of Marine Life's COMARGE project, the find was just one example of an unexpected array of marine biodiversity.

A PLACE IN THE SUN

Part of the space-shuttle programme's remit is to service the International Space Station, a commitment that will come to an end when the shuttles retire in 2010. The space station, docked with space shuttle *Endeavour*, is seen here transiting the Sun.

FROGFISH: D. HALL/SEAPHOTOS.COM; SEA CUCUMBER: L. MADIN, WOODS HOLE OCEANOGRAPHIC INST.; CAPSID: J. PAN & Y. J. TAO; MARMOSET: E. SASAKI; IDA: FRANZEN ET AL. PLOS ONE **4**, E5723 (2009); WORMS: K. J. OSBORN; TOAD: J. LARSEN MAHER/WCS

GREEN GENES

These glowing paws belong to some remarkable marmosets. The animals are the world's first transgenic primates — born to parents with altered genes that express green fluorescent protein. The achievement could help to produce primate models of human diseases.
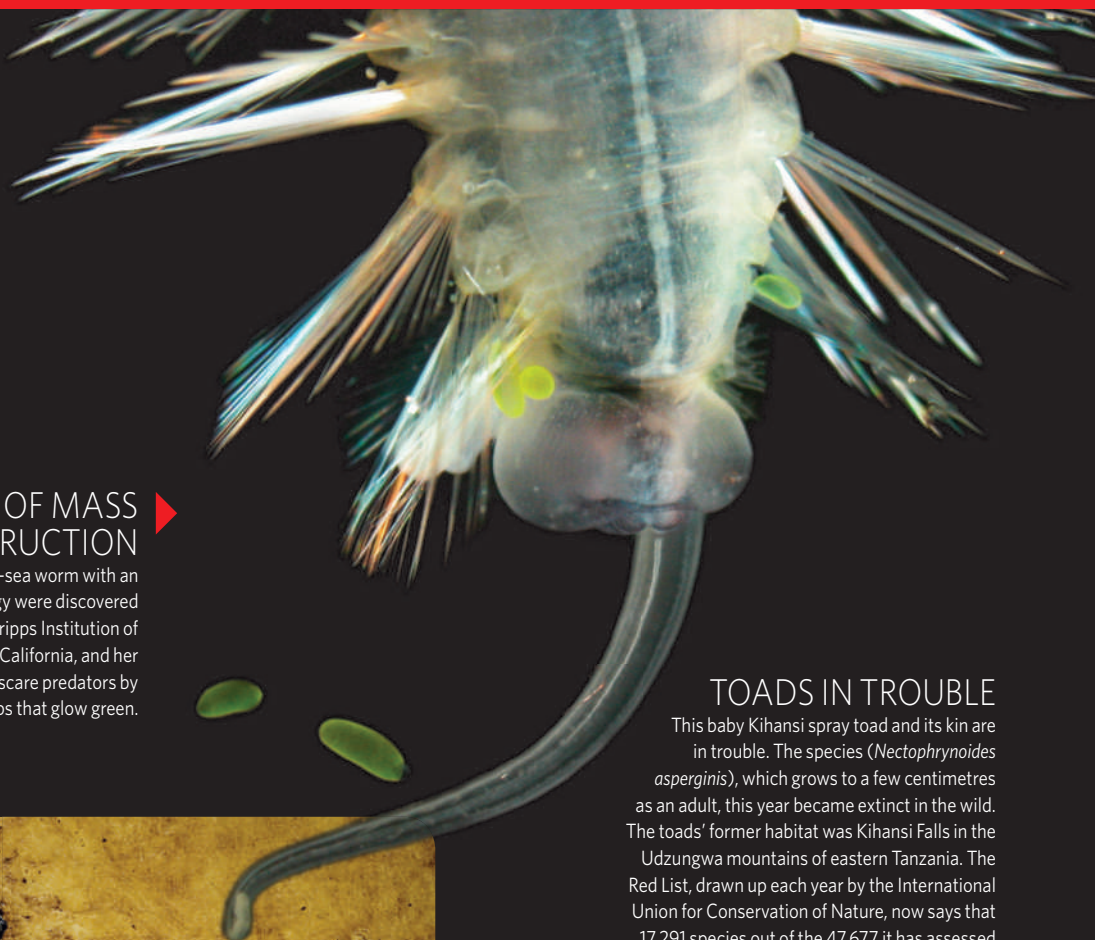
WORMS OF MASS DESTRUCTION

Several species of deep-sea worm with an unusual defensive strategy were discovered by Karen Osborn of the Scripps Institution of Oceanography in La Jolla, California, and her colleagues. The worms scare predators by releasing little bombs that glow green.

TOADS IN TROUBLE

This baby Kihansi spray toad and its kin are in trouble. The species (*Nectophrynoides asperginis*), which grows to a few centimetres as an adult, this year became extinct in the wild. The toads' former habitat was Kihansi Falls in the Udzungwa mountains of eastern Tanzania. The Red List, drawn up each year by the International Union for Conservation of Nature, now says that 17,291 species out of the 47,677 it has assessed are threatened with extinction.

WEAK LINK

In the year that celebrated 200 years since Charles Darwin's birth, finding a supposed 'missing link' to human evolution raised a media storm. The hype around Ida (*Darwinius masillae*), a 47-million-year-old primate-like fossil, soon turned sour — she is much closer to lemurs than humans on the tree of life.

THE POWER PLAYER

As a physicist, he found a way to capture atoms and won a Nobel prize. Now he is marshalling scientists and engineers to transform the world's biggest energy economy. **Eric Hand** profiles the US energy secretary, *Nature*'s Newsmaker of the Year.

STEVEN CHU is heading home on a bright day in October. His motorcade of government cars powers up the slope of Cyclotron Road, past the fragrant stands of eucalyptus and through the guard station at the entrance of Lawrence Berkeley National Laboratory. The vehicles continue along Chu Road and come to a stop near the top of the hill.

The man after whom the road is named heads into Building 50, which housed his office for the five years that he ran this laboratory overlooking the University of California, Berkeley. Inside an auditorium, 225 former colleagues await his arrival. Some wear suits; others slouch in hooded sweatshirts and sandals. There is an eager anticipation in the air, and moments before Chu arrives, the crowd grows quiet. Orange-vested security guards, armed with walkie-talkies, open the doors, and Chu walks down to the podium, his entourage trailing.

"It's very good to be back here," he says, flipping open his computer. "You people know I do my own PowerPoints. That has not changed." He launches headlong into a fast-paced and scattered talk that leaps across dozens of topics, all under the banner of climate change. He clicks ahead to the crucial slide — the one that shows actual measurements of rising global temperatures outpacing what would be expected without all the carbon dioxide that humans have spewed into the atmosphere. "Here's the evidence," he says. "I have to play this over and over again."

Such is his task back in Washington DC, where Chu now works as Secretary of the Department of Energy (DOE) and a member of President Barack Obama's cabinet — the first Nobel-prizewinning scientist to hold such a high office in the US government.

He is charged with transforming the world's biggest energy economy, and he has assumed the role of persuader-in-chief, trotting before Congress to explain the science of climate change and his plans for combating it. Meeting regularly with representatives and senators, he targets sceptics and walks them through the data. "I say, 'Come to my office and we'll talk about it'," he explains. "At the very least you can put a little doubt in their minds. If they're so sure it's natural causes, they may be less sure." It helps to have a Nobel prize, he adds.

In confronting what he sees as the most pressing problem facing the world today, Chu looks back in time to chart a way forwards. The Berkeley lab he once ran is the descendant of the Radiation Laboratory, where the physicist Ernest Lawrence helped find ways to enrich uranium for the Manhattan Project. Chemist Glenn Seaborg's team discovered plutonium there, and theoretical physicist Robert Oppenheimer worked just down the hill before heading into the New Mexico mountains to build the first nuclear bombs.

Chu plans to tackle climate change by reviving the scientific and technological urgency of the Manhattan Project — enlisting some of the nation's best minds to find a way to power the world without ruining it. His plans start at home, where he is trying to push the ponderous DOE to support riskier research that could yield huge dividends.

With a budget of US$27 billion, the department runs 17 national laboratories, oversees America's nuclear stockpile and manages the environmental clean-up after the early nuclear age. It is the largest source of funds for physical-science research in the United States, and this year Chu had a much bigger pot to dole out. Just one month into his tenure, Congress gave the agency $37 billion in economic stimulus money — funds that Chu is steering towards renewable energy, nuclear power, carbon-sequestration pilot plants and projects to modernize the electric grid, all of which should help to solve the climate problem. "They say that necessity is the mother of invention and this is the mother of all necessities," he says. "So we're going to get the mother of all inventions. And it's not going to be just one, it has to be many."

Hands-on manager

In the 1980s, Chu made his name scientifically by trapping atoms using lasers tuned with the utmost precision. Now he is applying that same mastery of detail to a vastly more complex system: an agency of 100,000 people working on all aspects of energy and nuclear issues.

Some Washington veterans have questioned whether Chu's research talent and hands-on style of management will serve him well, both at the DOE and amid the harsh political environment of the nation's capital. He has made some mistakes, notably in his dealings with Congress. But nearly a year into his tenure, Chu has proved that he is a quick learner. He has established himself as a voice that can be trusted by politicians of various stripes. He has helped to bridge international divides, particularly between the United States and China. And he has lured some top scientists from industry and universities to join him at the DOE in his quest.

Carol Browner, Obama's climate tsar, works often with Chu as part of the president's 'green cabinet', a group of senior officials who oversee environmental matters. "I think he's going to turn out to be the best energy secretary ever," she says. Praise also flows from some Republican politicians. Samuel Bodman, who led the DOE for former president George W. Bush, says that Chu has "shown skills as a manager. I think it was an inspired choice by the president to pick him."

Growing up in a New York suburb during the 1950s, Chu and his two brothers learned quickly that academic

> **"Necessity is the mother of invention and this is the mother of all necessities."**
> — Steven Chu

excellence — and competition — were family traditions. The boys would watch *College Bowl*, a 1960s television quiz show, and "the three of us would shout out answers and try to beat the contestants", recalls Morgan Chu, the youngest brother and a high-profile lawyer in California.

Chu's father and mother fled China during the Second World War and both did graduate work at the Massachusetts Institute of Technology (MIT) in Cambridge. The eldest son, Gilbert, followed the path of academic prestige — accumulating science degrees from Princeton University in New Jersey and MIT before gaining an MD from Harvard University in Cambridge, Massachusetts. Morgan did a PhD in social science before heading to Harvard Law School. Steven, on the other hand, was the A-minus student who favoured tinkering over schoolwork. In a family of Ivy Leaguers, he says he was the "academic black sheep", who settled for the University of Rochester in New York, where he studied mathematics and physics. Family pressures, he says, drove him — and frustrated him — early on, but once at Rochester, his facility for science flourished. "All of a sudden, the things they wanted me to do were very natural," he says.

On entering graduate school at Berkeley in 1970, Chu began a love affair with lasers. The work that was once a chore became the focus of an obsessive energy. "I've never been that good at apportioning time," he says. "When I got really excited about something, I would dig into it. It turns out that is a quality that the best researchers have." Another Berkeley graduate student, Phil Bucksbaum,

recalled nearly getting into a fist fight with Chu because he was being "bossy about the lasers", until a third student, who had studied with Chu at Rochester, explained to Bucksbaum: "It's the way he always has been. Focused and brusque," says Bucksbaum.

Chu's graduate work using polarized light to probe atomic transitions was good enough for him to get a job at Bell Labs in New Jersey, then a utopia for basic research. Chu thrived there, but he also made sacrifices. As his work progressed, he spent more time away from home, says his ex-wife, Lisa Chu-Thielbar. Sometimes, she would smuggle his first son, Geoffrey, under her overcoat onto the laboratory campus to catch some time with his father. "He was always a scientist first and a father second," says Chu's second son, Michael, who doesn't fault his father for the singular focus that allowed him to achieve so much. "The ambition was all intellectual and scientific. Steve never cared about money. He didn't even care about advancement," says Chu-Thielbar.

After seven years at Bell Labs, Chu had a key insight in 1985 into how to trap atoms. He crossed six lasers to form what he called "optical molasses", a goo of photons. It slowed atoms nearly to a standstill, making them sluggish enough to be held by the electromagnetic forces of an additional laser.

A year later, in the winter of 1986, Chu glimpsed the foundation of his Nobel prize through the windows of a vacuum chamber. Sodium atoms, cooled in optical molasses to 240 millionths of a degree above absolute zero, grew

bright orange as they fell, one by one, into a trap the size of a sand grain. A colour photo, the first ever published in *Physical Review Letters,* provided the proof of his success (S. Chu, J. E. Bjorkholm, A. Ashkin and A. Cable *Phys. Rev. Lett.* **57**, 314–317; 1986). The work would spawn applications across several disciplines. It provided biologists with 'optical tweezers' — ways to manipulate individual biomolecules, such as DNA. And it gave other atomic scientists the tools to create Bose–Einstein condensates, the super-cooled states of matter that can trap light and bring photons to a standstill, in a reversal of Chu's original technique.

By 1987, Chu was ready to move back to academia. He had offers at Harvard and Berkeley, but was intrigued at the idea of helping to build up a less-celebrated physics department at Stanford University in Palo Alto, California. It was a good plan. Stanford soon became a powerhouse; beginning in 1995, physicists there would win four Nobel prizes in a row, including Chu in 1997. While at Stanford, Chu started to push off in new directions, personally and professionally. He divorced Chu-Thielbar and married Jean Fetter, a physicist and former dean of admissions at Stanford. He took on graduate students with interests in biology and helped to convince the Stanford administration to build a $150-million biophysics centre.

But in 2004, just after that centre was completed, the Lawrence Berkeley National Laboratory (LBNL) came

> "Steve never cared about money. He didn't even care about advancement."
> — Lisa Chu-Thielbar

calling. Chu, who had never managed anything bigger than a physics department, was ready to make the leap to running the laboratory, which now has 4,000 employees and a $650-million budget. He showed his mettle early on, pushing the University of California system, which manages the LBNL, to use its debt service in an unprecedented way to finance new buildings for the lab, and fighting to save employee pension plans. Chu personally argued on behalf of his employees with the president of the University of California system until he relented, says Graham Fleming, a chemist at Berkeley and Chu's deputy at the time. "If one argument didn't work, he'd try another," he adds.
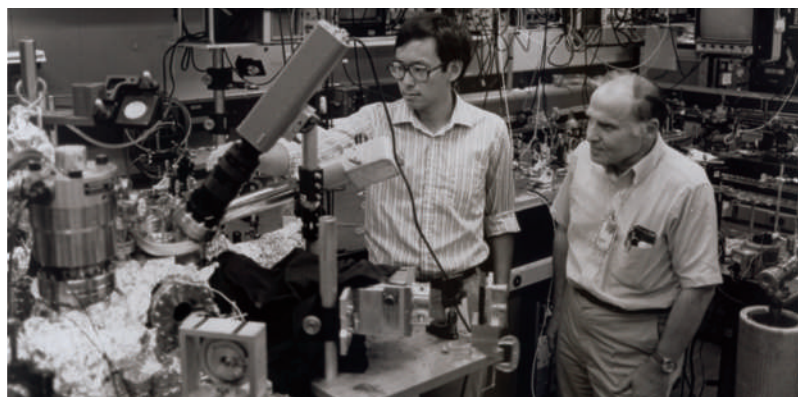
The climate crusader

Chu says that there was no one moment when he decided to devote himself full time to climate and energy puzzles. He had been digesting the science for years, reading reports of the Intergovernmental Panel on Climate Change. And he had pursued energy efficiency in his own life with his customary precision, complaining when workers skimped on insulation in his Stanford home. But soon after arriving at the LBNL, he decided that the time was ripe to resurrect an energy-research programme that had lain largely dormant since the fuel crisis of the 1970s. The lab was ready to revive those efforts, but it needed Chu's energy and vision, says Paul Alivisatos, who succeeded Chu as the head of the LBNL. "It's a bit like a supersaturated solution that you drop a seed crystal into," he says. "Steve was the seed crystal."

Chu gave the sprawling lab a purpose and convinced many scientists to make the switch, as he had, into energy research. He attracted large infusions of funding from the DOE and from the energy company BP. The lab launched major initiatives in biofuels and photovoltaics, but Chu also got involved in the little stuff. Alivisatos recalls Chu's interest in revamping a system of old, lumbering shuttle buses that circle the heights of the Berkeley Hills. Chu would stand on the balcony of the director's office and keep tallies of riders at a bus stop. "He thinks at an incredibly high level, but he also delves down into the finest detail," says Alivisatos. "And one of his abilities is to find the salient detail that matters enormously to the big picture and to show you how those connect."

But some thought that Chu went too far by micromanaging lab operations. "He doesn't see the necessity to get other people involved," says one scientist who knows him well but did not want to be identified as criticizing an official who controls so much research funding. "His whole career has been founded on his fantastic ability to worry about all the details himself. And that makes it hard for him to empower an effective staff."

Obama's election, and his campaign pledges to revamp the US energy system, created new opportunities for Chu. A few weeks after the election, Chu flew to Chicago to meet with the president-elect. "A lot of people are telling me you're the person for the DOE," said Obama, according to Chu. Rarely at a loss for words, Chu could only think to quip, "Who are these former friends of mine?"

Inspired by a sense of service, Chu planned to accept the job, but he did have a demand. He had seen the energy department hamstrung in the past by ineffectual people placed in posts to satisfy political obligations, so he

L. DEMATTEIS/REUTERS; A. WONG/GETTY

DOE

Clockwise from top left: the Nobel call, biking to work this year and time at Bell Labs.

wanted control over senior appointments. "There was a reasonable shot I could attract the right people. There are a whole bunch of people that have to lift this load," Chu says. Obama agreed, and Chu has recruited top talent such as Steven Koonin, former chief scientist of BP and provost of the California Institute of Technology in Pasadena, who is now undersecretary for science.

On the top floor of the energy department, portraits of secretaries past preside over the long, carpeted hallway leading from the elevators to the secretary's office. Most are of career politicians, with a few exceptions: Charles Duncan, who ran his family's coffee company, Donald Hodel, who would go on to lead two Christian evangelical groups, and James Edwards, a dentist. Bodman, Chu's predecessor, has an engineering degree from MIT. But Chu is the first scientist to lead an agency that has such an important role in physical-science research.

On an end table in his waiting room lie some recent biophysics papers on which Chu is a co-author. Chu puts the papers out to make a point — to visitors and himself — that he is still a working scientist. During his time at the LBNL, he kept a small research group of Stanford and Berkeley students, holding group meetings on Friday nights and on weekends. Even now, he says, he finds a little time for research during plane flights. Although Chu's October visit to the LBNL was his first lab-wide talk, he had visited before to check in with postdocs and meet new, young scientists — trips that Alivisatos calls Chu's "science vacations".

During an interview at his office, Chu settles into the centre of a couch, his back to an expansive view of Washington DC's mall and the Smithsonian Castle. At 61, Chu is resolutely trim. Although he no longer commutes to work by bike, as he often did at Berkeley, he manages long weekend rides and regularly climbs the seven flights of stairs to his office. Chu leans back when he listens, which is often, and leans forward when making a point. He is quick to crack a joke and eager to please. At least, he's that way with politicians (and reporters). With scientists, he can be impatient. "He does not suffer fools," says Michael Levi, an astrophysicist at the LBNL.

Chu was sworn in the day after President Obama.

Chu retains a scientist's candour — and that can sometimes get him into trouble. At his confirmation hearing, some senators jumped on Chu for calling coal "his worst nightmare" in a 2007 talk. (Chu says that the United States, China and India are unlikely to turn their backs on their huge coal reserves and that underscores the need to find clean ways to use the fuel.) A month after taking office, Chu slipped when he told reporters that it was "not in his domain" whether the Organization of the Petroleum Exporting Countries (OPEC) should cut oil production, an impolitic statement. He acknowledges that he was surprised at how his words have been magnified by the press.

Yet his inability to mince words is also an asset, especially in the floors below him in the DOE's headquarters, a fortress on concrete stilts. The DOE national labs have been characterized as inefficient, but that is in part because past safety and security lapses have led to a culture that stresses caution over aggressive research. When, during his talk at the LBNL, Chu mentions his desire to return to the original spirit of the labs, GOCO — government owned, but contractor operated — he gets a hearty round of applause. Chu says that the risk-averse culture, at both headquarters and the labs, must be changed. "The best way to protect yourself from something bad happening is to not do much."

Chu has already made headway. When he found out that billions of dollars in loans for energy projects that had been authorized in 2005 had not progressed, he insisted that they be pushed out in months, with the first one going to a solar-power company. It has helped to get involved personally, he says. Before closing on a $5.9-billion loan with Ford Motors, Chu says he was talking to the firm's chief executive every third day — an example that sent a clear message to his subordinates to act. "In certain areas, I'm not going away," he says. "The pressure is not going to let up."

To encourage more adventurous research, he has pushed to develop the Advanced Research Projects Agency-Energy, known as ARPA-E, which draws its inspiration from DARPA, the celebrated research programme run by

Blue sky, green tech

Under the leadership of Steven Chu, the US Department of Energy has started to fund high-risk energy projects through its Advanced Research Projects Agency-Energy. In October, 37 projects won a total of $151 million, including the following examples.

Carbon capture with artificial enzymes
Significant energy is needed to capture and store carbon dioxide from the combustion of fossil fuels. But a low-energy solution exists — in the human body. CO_2 from cells dissolves in the bloodstream then comes out of solution in the lungs. The enzyme involved, carbonic anhydrase, works at body temperature in both reaction directions. A US$2.3-million grant went to United Technologies Research Center in East Hartford, Connecticut, to explore the enzyme as a model for developing an artificial carbon-capture enzyme.

Giant liquid-metal batteries
One idea for improving batteries draws its inspiration from aluminium-smelting plants, which use huge, sustained, liquid reactions that suck up currents of half a million amps. A team from the Massachusetts Institute of Technology in Cambridge won a $6.9-million grant to try to reverse this process to create a giant battery. The idea could reduce energy-storage costs by an order of magnitude and scale up batteries to the size of buildings. These could store wind or solar energy during the day and release it at night.

LAWRENCE BERKELEY NATL LAB

Berkeley giants such as Ernest Lawrence (left), Glenn Seaborg and Robert Oppenheimer (right) have inspired Chu.

the Department of Defense that had an important role in creating the Internet. ARPA-E is designed to pursue high-risk, high-reward research on new forms of energy and conservation (see 'Blue sky, green tech'). The programme preceded Chu, but it's a pet of his, not least because he recommended it as a co-author on an influential study by the National Academies entitled *Rising Above the Gathering Storm*, which in 2005 warned of declining American competitiveness.

The ARPA-E concept will work, says Chu, only if the smartest reviewers are enlisted to pick out the most innovative ideas — otherwise incremental research is rewarded. "I unfortunately can't review all of the proposals myself," he told a group of clean-energy businesspeople in October, only half-jokingly. So Chu wrote a letter to the presidents of top research universities asking them to nominate their best researchers as ARPA-E reviewers. Five hundred responded to the call for duty.

Chu himself spent about two hours with the final set of proposals. Fleming, Chu's former LBNL deputy, says this sort of task suits his old boss. "I've never known anyone able to go away and come back 10 minutes later knowing so much about a new topic." And on the day that he visited the LBNL, Chu announced the 37 winning proposals, which would use $151 million of an initial $400 million given to the programme.

Chu's most ambitious idea has been to create eight focused and independent energy labs, modelled after the Manhattan Project, to develop technologies such as next-generation batteries and advanced nuclear power (see 'Chu's innovation factories'). But this is where he has run into the most trouble, and it exposes the limitations of the do-it-all-yourself approach. As Congress debated whether to fund Chu's new labs in fiscal year 2010, staffers found that they couldn't get the details on what, exactly, the DOE wanted. Would they be virtual labs, or permanent facilities? How many years would they be funded for? What mix of basic and applied science would be supported? "The hubs were just dropped on Congress," says one congressional staffer who adds that Chu's office did not provide consistent or timely information.

The communication problems with the Hill were on display at a hearing of a congressional appropriations committee in May. Senator Diane Feinstein (Democrat, California), a friend of Chu's, had a complaint. She had wanted to talk privately with him about some solar projects, but she had not been able to make an appointment to see Chu via his staff. "I'm a little bit surprised if you asked to see me and my staff said no," Chu replied.

"We just haven't gotten a response, that's sort of the way it's done," said Feinstein in an apparent attempt to educate the secretary on Washington customs. But Chu, who likes to deal with issues himself, did not seem to understand. "I'm still surprised," he said. "You actually have my private number."

In the end, when Congress doled out money to the DOE, Chu lost some battles. Money that he had proposed cutting from hydrogen research was reinstated. A $115-million education programme he had championed received nothing. Worst of all, for Chu, only three of his eight energy hubs were funded.

Chu's critics say that more attention to Congress could have alleviated the problems, but nearly a year into his tenure, he has not appointed an assistant secretary to head up his legislative-affairs office. Chu says the vacant position was not the problem. The issue was that he hadn't followed through himself. "The failure was on my part," he says, "because I wasn't communicating what the real vision was."

Energy ambassador

On a cold day in early December, Chu was preparing to travel to the United Nations' climate-change conference in Copenhagen. Before the trip, one of the last public events on his schedule was to appear with Secretary of Com-

Chu's innovation factories

Drawing on his experience at Bell Labs and his knowledge of the Manhattan Project, Secretary of Energy Steven Chu has proposed eight 'energy innovation hubs', in which scientists and engineers would work under one roof and be led by a strong director. Each hub would get $25 million a year for 5 years. This year, Congress funded just the first three in the list below.

Topic	Aim
Fuels from sunlight	Invent artificial photosynthetic systems to make liquid fuels directly from atmospheric carbon dioxide.
Energy-efficient building systems design	Create building-control systems analogous to the computerized optimization of car engines.
Modelling and simulation for nuclear reactors	Design fourth-generation nuclear reactors to go far beyond current designs.
Extreme materials for nuclear energy	Develop advanced materials for use in nuclear technologies, including fuels, shielding and waste.
Batteries and energy storage	Research new materials and structures to improve storage density and charging-cycle lifetimes.
Solar electricity	Find new photovoltaic or solar thermal approaches for generating electricity.
Carbon capture and storage	Reduce the cost and energy drain of carbon capture with novel absorbents; find new approaches to chemical, physical and biological separation.
Grid materials, devices and systems	Develop advanced materials for transmitting power; make 'smart' sensors that direct energy more efficiently.

Busy schedule: Chu likes to take care of many details on his own.

merce, Gary Locke, to talk about speeding up the process for granting patents on green technologies.

In July, the two secretaries went to Beijing together to meet with Chinese energy ministers. Locke, a prominent politician of Chinese descent, was greeted warmly. But Chu, with his Nobel-prize pedigree, was a rock star in a culture that reveres education. "He was like a Michael Jordan," says an administration official. "Everybody knew this guy."

Chu has taken a particular interest in China not because of his ancestry, he says, but because it emits more carbon dioxide than any other nation and it is also spending billions of dollars on clean-energy research. During the trip, Chu and Locke announced that the United States and China would jointly pursue research in areas such as energy efficiency and capturing carbon dioxide from coal-plant exhaust.

In his trip to Denmark, Chu reprised his role as energy ambassador. He announced plans to hold a conference next year with foreign energy ministers and pledged $85 million in US aid for renewable-energy projects in the developing world. For Chu, the summit served as a prelude to the fight next year, when he will use his main weapons — knowledge and powers of persuasion — to try to convince members of Congress to vote for a climate bill that would for the first time cap US emissions of greenhouse gases.

Chu says that when he ends his time as energy secretary, he will measure his success by two criteria: whether he aided adoption of a climate bill, and how much he changed the way that the DOE supports science. Those metrics would have seemed odd to a young scientist at

> "One of his abilities is to find the salient detail that matters enormously in the big picture." — Paul Alivisatos

Bell Labs in the 1980s who spent his days fretting over the precision of laser beams. Chu didn't plan on working his way to the upper echelons of the US government, where he is the first scientist since the cold war to play such an active part. "It just sort of happened," he says. "I followed the path first from going and doing the science, to getting very concerned about some issues that affect us all as a society, to finally saying, I can't sit idly by and occasionally give a talk on this. I really have to get proactive and put my money where my mouth is and do a career shift because it is that important."

But looking back, it's possible that the call to public service may have been whispering to Chu even during his graduate-school days at Berkeley, where the memories of the war effort remained fresh in the physics department. When Chu briefly took up sculpting at Berkeley, he chose to make a bust of Oppenheimer, the physicist-turned-manager who oversaw all details of the Manhattan Project.

Chu is now looking to another Berkeley star for inspiration. Lately, he has been reading the journals of Seaborg, who led the war-time team racing to extract plutonium for a bomb. Seaborg recounts how his group required fast-working Geiger counters that were not available at the time. So he pushed his crew to invent the needed detectors. For Chu, that sense of urgency in the face of a great threat stands out in Seaborg's work: "He kept saying: 'This isn't university research. We've got to move much faster'." ∎

Eric Hand is a reporter for *Nature* based in Washington DC. See Editorial, page 957.

CORRESPONDENCE

Goodbye to Darwin from a contemporary with vision

One night some 40 years ago, I was working late and alone in the library at the Marine Biology Laboratory at Woods Hole (in those days, the library never really closed), searching for something in the 1882 volume of *Archiv für Protistenkunde*. As I opened it, out fell a folded page from the magazine *The Nation* (still publishing today), dated 27 April 1882.

The page, headed 'Charles Darwin', was his obituary. As far as I know, it has not been reprinted or indexed in, for example, the Darwin archive at http://darwin-online.org.uk. There is no indication of the author's identity, although it would be gratifying if it were his supporter and friend, the American botanist Asa Gray.

I was impressed by the prescient observations on Darwin. For example, the final paragraph points out "There can be little doubt that Mr. Darwin's name will go down in history as that of the greatest scientific inquirer and the most pregnant scientific thinker that has lived since Newton. Since the beginnings of modern learning, probably no single idea has wrought upon the minds of men with such rich and manifold results as the idea of 'natural selection'; and it is evident that what we have already seen is but an earnest of vastly more that is to come."

Richard Kool School of Environment and Sustainability, Royal Roads University, Victoria, British Columbia V9B 5Y2, Canada
e-mail: rick.kool@royalroads.ca

Global Darwin: long kept under wraps in Pakistan

Marwa Elshakry's Opinion article (*Nature* **461,** 1200–1201; 2009, and see go.nature.com/97zlyr) makes no mention of the conflict of Darwin's ideas with popular religious beliefs in some conservative societies across the eastern world. There, the writings and thoughts of intellectuals, however influential, are no match for traditional religion.

For example, in Pakistan it was not until 2002 that a chapter on evolution was included for the first time in a school textbook, as a result of the federal government's educational reforms. The earlier decades of attempts to suppress scientific ideas were certainly not "enchanting".

Elshakry makes reference to Muhammad Iqbal, the Muslim thinker and reformer from early last century. Although Iqbal sought to challenge the traditional interpretation of religious beliefs and to understand religious principles in light of modern scientific thought, he avoided any direct mention of evolution or natural selection in his Urdu and Farsi writings. This was not because he was unaware of Darwin's works, but probably because he realized his audience was not yet ready to appreciate the significance of these ideas. Given their background of widespread illiteracy and poverty, deep-rooted social and religious conservatism, and colonial rule, religion was these people's last hope — and it was not the time to take that hope away.

Saheeb Ahmed Kayani National University of Sciences and Technology, Islamabad-44000, Pakistan
e-mail: saheebk@ceme.nust.edu.pk

Global Darwin: ideas blurred in early eastern translations

The early diffusion of Darwin's ideas into China resulted in multiple interpretations, imperfect translations and unsatisfactory terminology, as James Pusey notes in his Opinion article (*Nature* **462,** 162–163; 2009). However, he inadvertently implies that it was the Chinese scholar Yan Fu who translated 'evolution' as *jinhualun*, which means 'theory of progressive change'.

In fact, the word *jinhualun* originated in Japan in the 1870s, gaining popularity in China only after appearing in Ma Junwu's later translation of Darwin's *The Origin of Species*.

Instead, Yan Fu coined the term *tianyanlun*. The Chinese words *tian* and *yan* are layered in meaning, with *tian* translatable as 'heaven' and *yan* as 'development' or 'performance', among other concepts. But most would agree that *tian* corresponds nicely to the English word 'nature', whereas *yan* in this context denotes 'evolution'.

So, 'the theory of natural evolution' was Yan's preferred translation — a much more agreeable term, whether or not its alternative meanings are included.

David Flannery Department of International Studies, Macquarie University, Sydney, New South Wales 2109, Australia
e-mail: david.flannery@students.mq.edu.au

Readers may comment on the Global Darwin series at go.nature.com/97zlyr.

Don't forget the artists when studying perception of art

Martin Kemp in Books & Arts (*Nature* **461,** 882–883; 2009) suggests using functional neuroimaging to study the viewing and reception of artworks. But such direct measures of brain activity allow only for correlations between brain responses and the task of the viewer.

Clinical neuropsychologists have already studied the consequences of brain damage on cognition more directly. Insight into neurocognitive factors underlying art-making has come from, for instance, the effects of dementia on the abstract expressionist William de Kooning (1904–97) and of stroke on the German artists Lovis Corinth (1858–1925) and Otto Dix (1891–1969). Some milder conditions can even enhance productivity and creativity. For example, the metaphysical art of Giorgio de Chirico (1888–1978) may have been inspired by migraine or epilepsy.

Kemp focuses entirely on the beholder, as though — to paraphrase the French philosopher Roland Barthes (*Aspen* **5–6**; 1967) — the birth of the viewer must be at the cost of the death of the artist. However, art historians and neuroscientists also need to take into account the maker and the making of artworks — a collaboration that is successfully being developed in the Swiss Artists-in-Labs programme (www.artistsinlabs.ch/english/index.htm).

Olaf Blanke, Luca Forcucci, Sebastian Dieguez Laboratory of Cognitive Neuroscience, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Station 19, 1015 Lausanne, Switzerland
e-mail: olaf.blanke@epfl.ch

How much are we willing to pay for a fossil?

The outrage expressed by Elwyn Simons and others over the sale of a 47-million-year-old fossil for an enormous sum (*Nature* **460,** 456; 2009) may not be altogether justified.

A fossil's intrinsic value relates to its preservation, rarity, scientific interest and completeness. Arguably, this may be reflected as a commercial value to museums or private collectors.

The price a fossil is likely to fetch is of considerable interest to modern Chinese farmers, for example, as it was to their European predecessors around Bolca in the Italian Alps and Messel in Germany. Their excavations are often crude and can damage the fossil, but they still expect payment from interested professionals.

Worldwide closure of the

"Kepler claimed that the star followed by the Magi was the equivalent of the *stella nova* of 1604–05." Martin Kemp, page 987

market would have the unwanted effect of causing illegal trafficking. Regulations regarding fossil finds are necessary.

In Italy, these are very strict: like archaeological objects, fossils belong to the state and cannot be sold. However, their monetary value is decided by professionals appointed by an archaeological board; a small percentage of the sum is then divided between the discoverer and the owner of the land where the fossil was found. This law has proved efficient in protecting our natural and cultural heritage.

Elisabetta Cioppi, Stefano Dominici Museo di Storia Naturale dell'Università di Firenze, Sezione di Geologia e Paleontologia, Via La Pira 4, 50121 Firenze, Italy
e-mail: elisabetta.cioppi@unifi.it

'Snow joke as festive season gives rise to a blizzard of fake flakes

Parts of the world are once again knee-deep in images of snow crystals for the Christmas and New Year festivities. Unfortunately, the grand diversity of naturally occurring snow crystals is commonly corrupted by incorrect 'designer' versions — as illustrated by the faux octagonal snowflakes depicted in a *Nature* online subscription advertisement and, ironically, captioned "…for anyone who loves science".

The snowflake's natural sixfold symmetry stems from the water molecules' hexagonal crystal lattice, held together by a hydrogen-bonding network and the structural form of lowest energy under the ambient cold conditions. This hexagonal shape has been known since at least 400 years ago, when the astronomer Johannes Kepler published a treatise on the subject *On the Six-cornered Snowflake (De nive sexangula* Tampach; 1611*),* as a new-year's gift to his patron — modern editions are still available.

Beautiful photographs abound,

including those taken by Vermont farmer Wilson A. Bentley starting in 1885 (W. A. Bentley & J. Humphreys *Snow Crystals* McGraw-Hill; 1931), or see www.snowcrystals.com. Why then do many artists invent their own physically unrealistic snow crystals?

We who enjoy both science and captivating design should aim to melt away all four-, five- or eight-cornered snow crystals from cards, children's books and advertisements, by enlightening those who unwittingly generate and distribute them. Let's welcome this as an opportunity to share a discussion about the true beauty of science over a mug of hot punch.

Thomas Koop Department of Chemistry, Bielefeld University, 33615 Bielefeld, Germany
e-mail: thomas.koop@uni-bielefeld.de

Rewilding can cause rather than solve ecological problems

Prehistoric-restoration schemes such as those described in your News Feature (*Nature* **462**, 30–32; 2009) are highly unusual. Introducing a mix of native and exotic ungulates into former agricultural land could constitute a risky conservation strategy.

Reintroduction of native species to portions of their former range from which they were

extirpated is a well-established conservation tool. But there are no scientific grounds for introducing animals such as elephants, camels, cheetahs and lions into novel environments. Numerous scientifically driven concerns bear on these maverick programmes, including adverse effects of alien species on the ecosystems they are meant to foster; importation of diseases that may leap to native species; escapes that lead to hybridization; and predators jumping fences to endanger livestock.

There are sociopolitical concerns too, such as plundering wildlife from countries and ecosystems where they are naturally found in order to stock game parks, and persuading a conservation-weary public to accept large charismatic exotics as substitutes for contemporary native species and ecosystems.

We therefore advocate a moratorium on importing non- indigenous megafauna into ecosystems. Ill-considered, poorly documented introductions cannot be trusted to turn back the ecological and evolutionary clocks on anthropogenic change.

Tim Caro Department of Wildlife, Fish and Conservation Biology, University of California, Davis, California 95616, USA
e-mail: tmcaro@ucdavis.edu
Paul Sherman Department of Neurobiology and Behavior, Cornell University, Ithaca, New York 14853, USA

A toast to Mendeleev, who merits more than periodic honour

Before the year is out, let's raise a glass to the great Russian chemist Dmitrii Mendeleev, to celebrate the 140th anniversary of his periodic table of the elements. Russia has commemorated this, and the 175th anniversary of Mendeleev's birth, with a postage stamp (pictured) and a two-rouble silver coin.

Mendeleev's outstanding achievement was to organize all the chemical knowledge of the day into a single table and to predict the existence of new elements such as scandium, gallium and germanium. His periodic table, published in 1869, contained empty spaces to accommodate these as-yet undiscovered chemical elements.

Mendeleev's periodic law and periodic table of the elements were welcomed by the world's scientific community, and yet he received scant recognition for his work during his lifetime. He was never awarded a Nobel prize, for example. And the third Tsar Alexander is said to have blocked Mendeleev's election as a full member of the Russian Academy of Sciences, although he was allowed to continue as a corresponding member.

However, Mendeleev has been recognized more recently. This year, the American Chemical Society celebrated his periodic table during its national chemistry week, with the theme 'Chemistry — it's elemental'.

In keeping with terms such as Newtonian mechanics, Darwinian theory, Mendelian genetics and Watson–Crick hydrogen bonding, should the world not honour Mendeleev by referring to his achievement as the 'Mendeleev periodic table of the elements'?

Renad I. Zhdanov Functional Genomics and Lipidomics Lab, Institute of General Pathology and Pathophysiology, 8 Baltiiskaya Street, Moscow 125315, Russia
e-mail: zrenad@gmail.com

BOOKS & ARTS

Timely tale of avian flu

A gripping account of the spread of the bird flu virus across southeast Asia holds lessons for the global management of pandemics such as swine flu, explains **Ab Osterhaus**.

The Fatal Strain: On the Trail of Avian Flu and the Coming Pandemic
by Alan Sipress
Viking: 2009. 400 pp. $27.95

As the world responds to the rapid spread of swine flu (H1N1), it reminds us how much we have learned from other feared influenza pandemics, including some that did not materialize. One such example is that of avian flu (H5N1), which led to the deaths of more than 250 people in southeast Asia from 2003 onwards. In *The Fatal Strain*, journalist Alan Sipress tells that story through gripping first-hand accounts from his travels to the epicentre of the outbreak, and offers lessons for the management of future global influenza pandemics.

Sipress spoke extensively to patients infected with bird flu, their family members, doctors, scientists and national and international public-health officials. After explaining the origin of the H5N1 virus in wild birds and poultry, he describes the first fatal infection in a young boy in Hong Kong in 1997, and the identification of the virus by local and Dutch scientists. The

infection of a further 17 people in Hong Kong, of which 5 died, triggered a heroic public-health intervention by the Hong Kong authorities, advised by the World Health Organization (WHO). The culling of all the birds in the live-bird markets led to the immediate interruption of bird-to-human transmission. The virus's high fatality rate — more than 60% — prompted the WHO to advise national governments to prepare for the worst, although infection never reached pandemic levels.

Sipress points out differences between this H5N1 threat and the Severe Acute Respiratory Syndrome (SARS) epidemic — caused by the SARS coronavirus — which emerged at the same time as H5N1 and involved research by many of the same scientists. In the case of SARS, international collaboration coordinated by the WHO stopped the outbreak soon after it had started. But SARS was a relatively soft target compared with an influenza pandemic: in most cases, the SARS coronavirus spreads only after the third or fourth day of illness. By contrast, influenza can be transmitted when the first symptoms are not even apparent.

Sipress follows further human H5N1 cases in China, Vietnam, Indonesia and Cambodia,

including the few events in Indonesia in which human-to-human transmission was observed. He pays relatively little attention to the spread of the virus by wild birds and other means towards the Middle East, Europe and West Africa. Instead, he focuses on the underlying causes at the animal–human interface, highlighting important factors such as major changes in agricultural practices and mass production of poultry instead of the traditionally reared backyard flocks; the role of cock-fighting practices; and the consumption of duck-blood pudding and other raw poultry products. He also stresses the crucial role of Asia's live poultry markets, as first proposed by virologist and world-renowned influenza expert Rob Webster of St Jude Children's Research Hospital in Memphis, Tennessee.

Several political issues arose amid preparations for the expected avian flu pandemic. An intriguing example at the national level is the courageous role of another virologist, Yi Guan from Hong Kong University, who did part of his training with Webster. Guan took on Chinese authorities in his search for the origin of the virus in China's live-bird markets and wildlife, and did the same for the SARS coronavirus.

Sipress's book also highlights how the global prospect of a flu pandemic has deepened the divide between the rich and the poor. For example, the refusal of the Indonesian health minister, Siti Fadilah Supari, to collaborate and share further virus data with international and national organizations such as the WHO, the US Centers for Disease Control and Prevention or the US Naval Medical Research Center was based on a deep distrust that resulting vaccines or other intervention strategies would be available only to a limited number of wealthy nations. This rebellion spread from Indonesia to other countries such as Brazil, India and Thailand.

The Fatal Strain is a marvellous history of the involvement of individual people, professionals and organizations in the preparation for a major public-health disaster. Although Sipress is convinced that "the most important line of defence is not a lab or a hospital, but is the farmyard gate", scientists will enjoy it. ■

Ab Osterhaus is head of the Department of Virology at Erasmus Medical Centre, 3000 CA Rotterdam, The Netherlands.
e-mail: a.osterhaus@erasmusmc.nl

C. CORTES IV/REUTERS/CORBIS

Sleeping with the enemy: Asia's live-bird trade fuelled transmission of avian influenza to humans.

Johannes Kepler on Christmas

Kepler's interpretation of the supernova of 1604, *De Stella Nova*, interwove the science of astronomy with astrology and theology in an attempt to determine the correct birthdate of Jesus, explains Martin Kemp.

When was Jesus born? On 25 December in the year that starts the Christian calendar, is what wide belief would have it. Greek and Russian Orthodox churches retain the date of 7 January in accordance with the old Julian calendar, rather than the Gregorian revision that was introduced in 1582.

Internal evidence from the Bible rules out the designated year because Herod the Great, who was responsible for the massacre of the innocents, from which the infant Jesus escaped, is known to have died in 4 BC. This is one reason why the now-obscure Polish scholar, Laurence Suslyga, argued in his 1605 thesis on the birth and death of Christ that Jesus was actually born in 5 BC.

Suslyga's tract unexpectedly gained a significance beyond the specialist realm of biblical chronology. The circumstances are best explained by the great German astronomer, Johannes Kepler: "I found for sale at Graz a small book by Laurence Suslyga of Poland," with which Kepler agreed that at least "four years must be added to the Epoch of Christianity now in use". Suslyga provided Kepler with the necessary licence to correlate the dramatic appearance of a 'new star' in 1604 with the act of stellar navigation performed by the biblical Magi when they journeyed to see the Christ child in Bethlehem.

In 1604, astronomers keenly awaited the shift of the conjunction of Jupiter and Saturn into the 'Fiery Trigon' of the zodiac (Sagittarius, Aries and Leo), which initiated a new cycle of conjunctions in the Trigon — an event that was calculated to recur about every 800 years. Mars also moved close to Saturn and then Jupiter, thus outlining a triangular array that was of great interest to astrologers.

During his time in Prague as mathematician to the Holy Roman Emperor Rudolf II, Kepler observed this remarkable cluster of planets and the new star on 17 October 1604, when the clouds over the city finally lifted. The star burned brightly in the evening and was even visible as a morning star, located in the foot of the constellation of the Serpent Bearer, or Ophiuchus (pictured). He made his last observation of the star a year later, after

Artwork from *De Stella Nova* shows the constellation Ophiuchus with the 1604 supernova (N; in the foot, lower left) observed by Kepler.

which it faded from view as the flash from its explosion declined in intensity.

In 1606, Kepler published a pamphlet on the new star in German, while planning a more substantial account in Latin: *On the New Star in the Foot of the Ophiuchus and on the Fiery Trigon that Began Anew at its Rising. A Booklet Full of Astronomical, Physical, Metaphysical, Meteorological and Astrological Disputations* (beginning in Latin, *De Stella Nova in Pede Serpentarii …*). The second of two substantial appendices in the book dealt with the year of Christ's birth "in the light of the new pronouncements of Laurentius Suslyga".

In a series of arguments that even he admitted were hard to follow, Kepler claimed that the star followed by the Magi was the equivalent of the *stella nova* of 1604–5, and that it had arisen during a series of related planetary conjunctions in the years 7–5 BC — which he took to cover the period of Christ's conception and the Magi's journey to Bethlehem, as recounted in Matthew 2:9–10.

Kepler's treatise can be acclaimed as a work of science in the modern sense. For instance, he uses the lack of diurnal parallax from the new star — that is, the inability to measure its distance using the angle at two different points on Earth's surface — to infer that it resided in the sphere of the fixed stars. This provided further evidence that the sphere was not as immutable as the ancients had assumed. But *De Stella Nova*, as its extended title declares, explicitly involved metaphysics and astrology, and it climaxed in an appendix on biblical chronology.

Within this, we cannot separate what we regard as the real science from the astrology and theology. For Kepler, mathematics and meaning comprised a unified whole. The great astronomer, best known for identifying the elliptical paths of the planets, became increasingly concerned in the years following publication of *De Stella Nova* with the construction of a 'purified' version of ancient astrological wisdom, purged of the myths that had accumulated around the zodiacal signs. He later referred to the mystic symbolism of traditional astrology as "filthy mud" from which "one can glean even an occasional escargot, oysters or an eel for one's nutrition".

Kepler gained his "nutrition" from the mathematical ratios of the 'aspects' between the planets. He defined an aspect as "a geometrical construction [of an angle] between light beams of two planets here on Earth". The ratios were integral to the celestial geometry that manifests the mathematical "music" of the heavens. He explained that earthly nature cannot help but respond to the dictates of heavenly harmonies, and said that nature is affected by an aspect "just as a farmer is moved by music to dance".

De Stella Nova serves to remind us that it was not possible in the era of Kepler and Galileo to pursue astronomy in such a way that the mathematical study of the heavenly bodies was divorced from the theology of a heaven inhabited by God. ■

Martin Kemp is emeritus professor in history of art at the University of Oxford, Oxford, UK.

Personal favourites of 2009

Nature invited some of its regular reviewers to name their pick of this year's book publications.

PETER ANSTEY
The Fourth Part of the World
by Toby Lester
Free Press/Profile: 2009.

In this page-turning history of cartography, Toby Lester shows how maps themselves were the real treasure in humankind's quest to discover the shape, size and location of the habitable world, from ancient times to the age of Christopher Columbus and Amerigo Vespucci.

With a journalist's eye for detail and a historian's feel for accuracy, Lester shows how the construction of new maps and the rediscovery of old ones were as important as voyages of discovery by those such as Marco Polo. This is a gripping story of how the charting of the world involved science, politics, theology and courage, culminating in an age when solving the problem of longitude was still a pipe dream. It is a saga about maps, their style, their vision, their structure, their failings and, above all, their secrets.
Peter Anstey teaches early modern philosophy at the University of Otago, Dunedin, New Zealand.

FRANK CLOSE
The Strangest Man: The Hidden Life of Paul Dirac
by Graham Farmelo
Faber & Faber: 2009.

Among scientists, Paul Dirac is regarded as being in the same league as Albert Einstein. Dirac is renowned for his 1928 prediction of antimatter, as a result of combining Einstein's theory of special relativity with the new quantum mechanics. As this excellent biography by Graham Farmelo shows, Dirac's contributions to science were profound and far-ranging; modern ideas that have their origins in quantum electrodynamics are inspired by his insight. Yet in his home town he was overshadowed by fellow student Archie Leach — film star Cary Grant. On asking at the Bristol Records Office for material about Dirac, Farmelo received the response: "Who?"

Hopefully, *The Strangest Man* has brought greater public awareness of this remarkable scientist. The effortless writing style shows that it is possible to describe profound ideas without compromising scientific integrity or readability. Farmelo also sheds considerable light on Dirac's personality and the circumstances behind it. Several members of the physicist's extended family had acute depression, and Farmelo makes a strong case that Dirac was autistic. The book makes one realize that great gifts sometimes come at great personal cost.
Frank Close is professor of physics and a fellow at Exeter College, University of Oxford, Oxford, UK.

W. F. BYNUM
The Oxford History of English Lexicography (Volumes I and II)
Edited by A. P. Cowie
Oxford Univ. Press: 2009.

Indisputably Charles Darwin's year, 2009 was also the 300th anniversary of the birth of Samuel Johnson (1709–84). His *Dictionary of the English Language* (1755) — a milestone in the development of what is now the lingua franca of science — features centrally in my favourite book of this year, *The Oxford History of English Lexicography*. The essays in these two large volumes include pieces on the evolution of specialized dictionaries of science, technology and medicine.

Cowie's authors remind us of the power of correctly used language in communication. Johnson knew many physicians and natural philosophers, so was aware that science deals with nature's material realities. His own job engaged only with the words they used: "Words are the daughters of the earth … things are the sons of heaven." But, he added significantly, language is "the instrument of science".
W. F. Bynum is professor emeritus of the history of medicine at University College London, UK.

GEORGINA FERRY
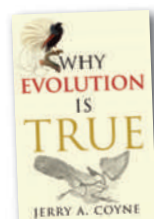**Francis Crick: Hunter of Life's Secrets**
by Robert Olby
CSHL Press: 2009.

Francis Crick was possibly the greatest biologist of the twentieth century — not bad going for a physicist. Robert Olby's biography is the closest we are likely to get to a definitive account of Crick's life, yet in September it slipped onto the shelves almost unnoticed.

Meticulously documented and well-written, *Francis Crick* retells the story of the discovery of the DNA structure and the cracking of the genetic code. But what did Crick do next? Less well known is his 25-year quest for that elusive prey, the neural basis of consciousness.

Olby knew Crick for almost 40 years, and the book shows respect and admiration for its subject. But it is also dispassionate in revealing a man of such breathtaking intellectual arrogance that he could dismiss much of philosophy as barren and argue for eugenic policies to control the fertility of the working classes. A useful reminder that rarefied intellectual heights are not necessarily the best vantage point from which to survey the rich landscape of human society.
Georgina Ferry is a writer based in Oxford, UK.

XU XING
Why Evolution is True
by Jerry A. Coyne
Oxford Univ. Press: 2009.

Jerry Coyne's book *Why Evolution is True* demonstrates the validity of evolutionary theory using the latest available data. Coyne is hugely knowledgeable, and an excellent writer. He uses multiple strands of data in such a clear and sometimes humorous way that an open-minded reader can hardly avoid learning a great deal about evolution: from feathered dinosaurs to vestigial organs, and from the unexpressed or fragmented genes present in all life forms to the complex distribution of Earth's flora and fauna.
Xu Xing is a professor at the Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China.

ELENA CATTANEO
Perché gli scienziati non sono pericolosi (Why Scientists Are Not Dangerous)
by Gilberto Corbellini
Longanesi: 2009.

Medical historian Gilberto Corbellini argues that Italy is, metaphorically, infected with an anti-science virus. He contends that sectors of the political and intellectual establishments believe that society would be better off without scientists. He denounces attempts by the Catholic Church and other forces to limit scientists' freedom to do research. To counter this worrying trend, Corbellini traces the steps that lead to a collective strength based on science and a full embrace of democracy. The book concludes that there is only one condition in which scientists are dangerous: when they stop being scientists, renounce intellectual independence and bow to politics and local lobbies.
Elena Cattaneo is director of the Centre for Stem Cell Research, Department of Pharmacological Sciences, University of Milan, Italy.

NEWS & VIEWS

AGEING

Diet and longevity in the balance

Thomas Flatt

Dietary restriction promotes longevity but impairs fecundity in many organisms. When the amino acids in a diet are fine-tuned, however, lifespan can be increased without loss of fecundity — at least in fruitflies.

It's common wisdom that eating the right food, and not too much of it, is good for you. Dietary restriction — reduced food intake without malnutrition — indeed prolongs lifespan in organisms ranging from yeast, worms and flies to rodents, monkeys and possibly humans. But dietary restriction also has its costs: it often impairs fecundity, possibly because maintenance of the soma (the non-germline parts of an organism), and thus long life, are incompatible with the metabolic demands of reproduction in such circumstances. Biologists have long thought that an organism's response to food shortage is an evolutionary device that allows individuals to survive a famine by diverting resources away from reproduction and reallocating them to essential functions for survival[1,2].

On page 1061 of this issue, Grandison *et al.*[3] report that this idea is almost certainly wrong. They find that dietary amino acids are responsible for shortening lifespan and increasing reproduction in the fruitfly *Drosophila melanogaster*, but that both longevity and fecundity can be maximized when intake of these nutrients is finely tuned.

Great strides have recently been made in our understanding of dietary and caloric restriction[4–8]. It has become clear that rich diets shorten life, not because of excess calories but rather because of dietary imbalance, with lifespan and fecundity being maximized at different nutritional optima[5]. Specific nutrients are implicated in dietary restriction, especially amino acids, the building blocks of proteins. For example, reducing the intake of casein, a major amino-acid source, extends lifespan but decreases fecundity in *Drosophila*[6]. Similarly, methionine restriction promotes longevity in flies, rats and mice[7,8]. Although these studies suggest that amino-acid deficiency might account for the effects of dietary restriction, Grandison *et al.*[3] have gone several steps further.

The authors have identified the specific nutrients that modulate lifespan and reproduction by manipulating all of the components of a defined diet. In a series of painstaking experiments, they fed female flies a restricted diet that extends lifespan at the expense of

Figure 1 | Signalling pathways possibly influenced by dietary amino acids in *Drosophila*. Grandison *et al.*[3] show that the lifespan-increasing and fecundity-reducing effects of dietary restriction can be attributed to amino-acid limitation. During full feeding, amino acids tend to shorten lifespan and promote egg production. These effects seem to operate via insulin-like signalling, mediated by insulin-like peptides produced in the brain, as mutant flies lacking insulin receptors are protected against the lifespan shortening and loss of fecundity that usually accompany full feeding. During development and growth, insulin signalling interacts with the nutrient-sensing TOR pathway. Because TOR signalling has been implicated in dietary restriction[4,9], insulin/TOR crosstalk could integrate the effects of dietary amino acids on lifespan and reproduction. Blue, fat body; green, digestive system; InR, insulin-like receptor; TOR, target of rapamycin.

fecundity, and then tried to restore the short-life and high-fecundity characteristics of fully fed flies by adding back specific nutrients. Adding carbohydrates, lipids or vitamins made no difference. But adding amino acids shortened lifespan and increased egg production to the level observed under full feeding. The question then was whether all amino acids were contributing equally to this astonishing effect.

Amino acids can be divided into two types: non-essential (those that can be synthesized by an organism) and essential (those that must be supplied in the diet). Grandison *et al.*[3]

found that adding all non-essential amino acids only marginally shortened lifespan and did not change fecundity, whereas adding all essential amino acids decreased lifespan and increased egg production as much as combining all amino acids or full feeding. The authors next manipulated each essential amino acid individually. Remarkably, methionine alone increased fecundity as much as full feeding but without reducing lifespan. In fact, no single amino acid shortened lifespan to the fully fed state. Yet, adding all essential amino acids except methionine failed to reduce lifespan, suggesting that methionine together with one or several other essential amino acids is responsible for the lifespan-shortening effect of full feeding. This finding is consistent with work showing that methionine restriction promotes fly longevity[7].

So how does methionine, either by itself or in combination with other amino acids, influence lifespan[7,8]? One possibility is that methionine restriction counters oxidative damage: dietary methionine deficiency increases the levels of the antioxidant glutathione in rats and protects mice from oxidative liver-cell injury[8]. Another explanation might be that methionine restriction extends lifespan by reducing signalling through a major regulator of longevity in worms, flies and rodents[2,4,7,8], the IIS — insulin/insulin-like growth factor (IGF) — pathway. Indeed, methionine-deficient mice show lowered levels of IGF-1 and insulin[8]. This is particularly interesting in the light of Grandison and colleagues' results: adding back essential amino acids to the diet decreased lifespan strongly in normal flies, but only slightly in mutant flies lacking the insulin-like receptor that mediates IIS activity. Addition of methionine did not promote fecundity in these mutants. These results imply that the IIS pathway mediates key effects of amino acids on ageing and reproduction (Fig. 1).

During development and growth, IIS is known to interact with 'target of rapamycin' (TOR) signalling, a nutrient-sensing pathway also implicated in dietary restriction[4,9], and it is tempting to speculate that amino acids act through TOR to affect lifespan (Fig. 1). In line with this idea, a repressor protein, 4E-BP,

which lies downstream of TOR, is required for lifespan extension in flies when their diet is restricted[9]. Furthermore, dietary amino acids stimulate insulin secretion in the brain through a TOR-dependent mechanism in the *Drosophila* fat body, the equivalent of mammalian liver and adipose tissue[10]. The connections between dietary amino acids, IIS and TOR clearly merit further investigation.

The findings of Grandison *et al.*[3] also have implications for understanding the trade-off between survival and reproduction. Although reproduction shortens lifespan in various organisms[2,11], the mechanisms involved have remained elusive. The most widely accepted view is that when resources are invested in reproduction, they are unavailable for somatic maintenance and survival[1]. Alternatively, reproduction might inflict direct somatic damage independent of nutrient allocation[11]. Both explanations, however, seem to be at odds with the observations of Grandison *et al.* because methionine alone can restore full fecundity to long-lived flies on an otherwise restricted diet. If the diet is correctly balanced, both fecundity and lifespan can be maximized, without any apparent trade-off between them.

But are things really this simple? Probably not. Although dietary restriction can extend lifespan when flies or worms are sterilized[2], it does not further improve the lifespan of long-lived worms lacking germ cells[12]. Dietary restriction and reproduction might thus converge on the same mechanisms regulating lifespan[12]. That we don't yet fully understand the intricate connections between diet, metabolism, reproduction and lifespan is underscored by the finding that ablation of germline cells extends lifespan by decreasing fat storage in the worm *Caenorhabditis elegans*[13]. Nonetheless, if the present results[3] in the fly are generally applicable, even mammals might be able to enjoy a long life without loss of fecundity by virtue of a suitably balanced diet. ■

Thomas Flatt is in the Department of Biomedical Sciences, Institute of Population Genetics, University of Veterinary Medicine Vienna, Veterinärplatz 1, A-1210 Vienna, Austria.
e-mail: thomas.flatt@vetmeduni.ac.at

1. Kirkwood, T. B. L. *Nature* **270,** 301–304 (1977).
2. Flatt, T. & Schmidt, P. *Biochim. Biophys. Acta* **1790,** 951–962 (2009).
3. Grandison, R. C., Piper, M. D. W. & Partridge, L. *Nature* **462,** 1061–1064 (2009).
4. Mair, W. & Dillin, A. *Annu. Rev. Biochem.* **77,** 727–754 (2008).
5. Skorupa, D. A., Dervisefendic, A., Zwiener, J. & Pletcher, S. D. *Aging Cell* **7,** 478–490 (2008).
6. Min, K.-J. & Tatar, M. *Mech. Ageing Dev.* **127,** 643–646 (2006).
7. Troen, A. M. *et al. Age* **29,** 29–39 (2007).
8. Miller, R. A. *et al. Aging Cell* **4,** 119–125 (2005).
9. Zid, B. M. *et al. Cell* **139,** 149–160 (2009).
10. Géminard, C., Rulifson, E. J. & Léopold, P. *Cell Metab.* **10,** 199–207 (2009).
11. Tatar, M. & Carey, J. R. *Ecology* **76,** 2066–2073 (1995).
12. Crawford, D., Libina, N. & Kenyon, C. *Aging Cell* **6,** 715–721 (2007).
13. Wang, M. C., O'Rourke, E. J. & Ruvkun, G. *Science* **322,** 957–960 (2008).

APPLIED PHYSICS

Bubbly but quiet

Edwin L. Thomas

An array of air bubbles in a rubber-like material can be made to block the transmission of sound. This finding might help in the design of soundproof walls for music rooms and urban apartments.

The past decade has witnessed growing interest in the fabrication of artificial materials that can interact with waves. Following on from successful efforts to microfabricate photonic crystals, thereby producing novel methods for controlling electromagnetic waves, researchers have started looking for more effective ways to control mechanical waves such as sound waves. When we listen to a symphony, we are detecting acoustic waves in our frequency range of hearing (sonic waves) — from 20 hertz to about 20 kilohertz (middle C has a frequency of 261 Hz). Higher frequencies, spanning 20 kHz to 1 gigahertz, are termed ultrasonic waves, and are commonly used for applications in medical imaging. At higher frequencies still, from about 1 GHz to the terahertz regime, are hypersonic phonons, which carry heat. Writing in *Applied Physics Letters*, Leroy and colleagues[1] describe a new material construct for damping the transmission of ultrasonic phonons.

A photonic (for light), or equivalently phononic (for sound), crystal is an artificially engineered material made of a regularly repeating arrangement of elements that allow impinging waves to undergo destructive interference: waves with frequencies that fall within a range known as the bandgap are prevented from propagating in the material. For destructive interference to occur, the period of the structure must be of the order of the wavelength. This means that, to block sound for human hearing, the periods would be very large, and the overall structure quite unwieldy — metre-sized! In fact, sound attenuation has been serendipitously achieved on Eusebio Sempere's phononic sculpture *Órgano*, which is on display in the garden of the Juan March Foundation in Madrid (Fig. 1). This structure has a bandgap at around 1.6 kHz, and is made of 3-centimetre-diameter metal rods arranged on a 10-cm-period lattice, fixed on a circular platform of 4 metres diameter[2].

But there is another way to block sound: resonance. In 1933, Marcel Minnaert published[3] a paper in the *Philosophical Magazine*, called 'On musical air-bubbles and the sounds of running water', in which he outlined the basic theory of the origin of the sound emitted by an air bubble formed in water. His key insight was that the bubble undergoes radial oscillation (a 'breathing mode') inside the surrounding fluid. Minnaert found that the resonant frequency of the bubble (ω_0) depends on the bulk modulus of the air (β_{air}; a measure of a substance's resistance to compression), the density of the water (ρ_{water}) and the radius of the bubble

JUAN MARCH FOUNDATION

Figure 1 | Not just a piece of art. Eusebio Sempere's sculpture *Órgano*, on display at the Juan March Foundation in Madrid, is a fortuitous example of a phononic crystal[2], a periodic arrangement of structures — in this case, metal tubes — that can block sound waves.

(R_{bubble}): thus, $\omega_0 = (3\beta_{air}/\rho_{water})^{1/2}/R_{bubble}$.

Interestingly, the low bulk modulus of air (about 10,000 times smaller than that of a typical liquid) results in a low resonant frequency, one that is within the range of human hearing even for quite small bubbles (R_{bubble} of the order of a millimetre). If, instead of creating bubbles that emit sound, one has a liquid containing a host of randomly arranged bubbles, incident waves of the correct frequency could set each bubble into resonance, thus absorbing the waves' energy and limiting the propagation of sound in the medium. If one could make bubbles of uniform size and place them in a periodic array of the correct spacing, the material would have two independent mechanisms for deterring the propagation of waves: energy absorption through the single-bubble resonance and destructive interference from Bragg scattering of the waves, owing to the periodic bubble array. The trick is to somehow create and fix in place the right sort of bubble array in a liquid. But it is difficult to make bubbles that have uniform size and that, once made, will remain inside a liquid.

In their experiment, Leroy and colleagues[1] circumvent this problem by employing a rubber-like material — a soft, castable elastomer called polydimethylsiloxane (PDMS) — in place of the liquid. They used a hard template to create a two-dimensional periodic array of cavities (bubbles) in each of a number of layers of the elastomer, and then stacked the layers to form a three-dimensional array of cavities. The cavities were actually cylinders with an aspect ratio near one, not perfect spheres, but this doesn't affect things much. Because PDMS is not a true liquid, one needs to modify the expression for the Minnaert resonant frequency to take into account its non-zero shear modulus (the material's rigidity). Fortunately, PDMS has a shear modulus only about five times that of the bulk modulus of air, so that the resonant frequency of an air bubble in this soft solid remains low.

Because Leroy et al.[1] couldn't precisely align upper and lower bubble layers with one another, the overall stacked structure has some disorder. But it turns out that such stacking disorder doesn't alter the destructive Bragg interference when the incident waves are travelling along the direction perpendicular to the structure surface. By fabricating bubbles with diameters of about 80 micrometres, stacked in layers about 360 micrometres apart, the authors allow the two mechanisms to block the same waves. They find that sound is attenuated by a factor of 1,000 over a bandgap of 0.25 MHz. Their finding shows that, by combining resonance absorption and destructive Bragg scattering, a more compact structure could be used to block sound.

Thus, Leroy and colleagues' simple system paves the way for the development of more complex bubble-based phononic crystals that could yield a broader and deeper (greater attenuation) bandgap than the authors achieved. This could be accomplished by engineering the material

to have more than one structural period and bubbles of different sizes. Another possibility is to physically join individual bubbles; this would reduce the system's resonant frequencies and increase the effective damping of the waves and the width of the bandgap.

In future experiments, alignment of the elastomer layers could be obtained by various optical or physical means, providing Bragg interference for waves incident from various angles. Thus, by using different bubble sizes, connecting bubbles within individual layers, varying the spacing between layers and so on[4],

the soft, bubbly medium proposed by Leroy and colleagues may one day meet the challenges faced by garage bands trying to insulate their sound from their neighbours. ∎

Edwin L. Thomas is in the Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mail: elt@mit.edu

1. Leroy, V. et al. Appl. Phys. Lett. **95,** 171904 (2009).
2. Martínez-Sala, R. et al. Nature **378,** 241 (1995).
3. Minnaert, M. Phil. Mag. **16,** 235–248 (1933).
4. Feuillade, C. J. Acoust. Soc. Am. **98,** 1178–1190 (1995).

ASTROPHYSICS

Stellar revival in old clusters

Melvyn B. Davies

Observations of star clusters in the Milky Way show that collisions between stars as well as mass flow within binary systems can explain how the peculiar family of blue straggler stars came to be born.

What is the origin of the population of oddly massive stars that lurk in the cores of old star clusters? These stars have become known as blue stragglers because they tend to be bluer and younger than most of the clusters' stars. Their existence in star clusters is at odds with a simple picture of stellar evolution, which demands that they should have exhausted their nuclear fuel and evolved long ago to become cooling, Earth-sized stellar remnants known as white dwarfs. This fate awaits our own Sun, around five billion years from now. In this issue, Ferraro et al.[1] (page 1028) and Mathieu and Geller[2] (page 1032) present observations of star clusters that may provide important clues to the origin of blue stragglers.

The study of blue stragglers has a long history. By imaging a star cluster and plotting the brightness versus colour for the individual

stars, in what is known as a colour–magnitude diagram, one sees that the population of blue stragglers lies in an extension bluewards of the regular main-sequence band of stars on which most stars are located; main-sequence stars are powered by the fusion of hydrogen into helium in their cores. Blue stragglers were found in this way in the 1950s, for example in the globular cluster M 3 (ref. 3). Since then, observations have revealed blue stragglers in all rich star clusters.

Star clusters are hostile environments. The number density of stars in the centre of these systems is much higher than in most parts of the Galaxy. These regions are so crowded that stars will, over time, have close encounters with each other, and possibly even collide[4]. Computer modelling of stellar collisions shows that when two low-mass main-sequence
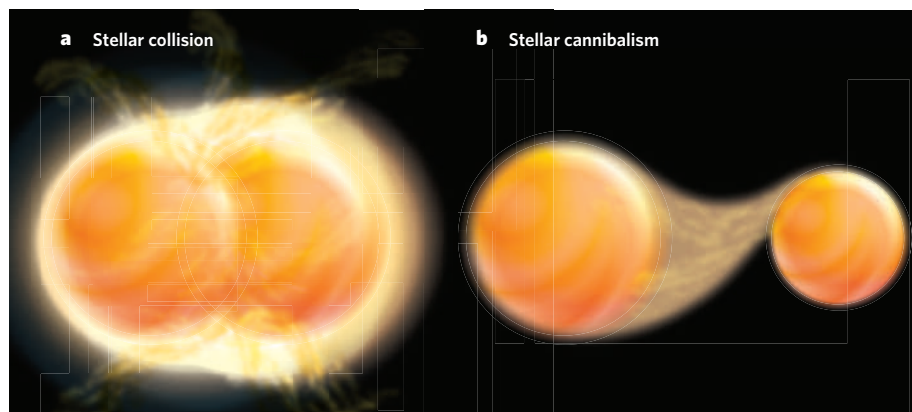

Figure 1 | Two ways to produce a blue-straggler star. According to theories of stellar evolution, the unusually massive stars dubbed blue stragglers that are seen in star clusters can be produced in two distinct ways: **a**, stellar collision, whereby the star is created through a collision between two stars; **b**, stellar cannibalism, whereby the star is formed via mass transfer from a companion in a binary star system. Observations of star clusters by Ferraro et al.[1] and Mathieu and Geller[2] confirm that both formation mechanisms are viable.

HANGOVERS
Uncongenial congeners

Whatever your tipple, too much of it and you'll suffer the next day. Yet it's commonly thought that dark-coloured alcoholic drinks such as bourbon produce worse hangovers than colourless alternatives such as vodka. Damaris Rohsenow and colleagues now provide experimental evidence of this (D. J. Rohsenow *et al. Alcoholism Clin. Exp. Res.* doi:10.1111/j.1530-0277.2009.01116.x; 2009). They also find that hangovers caused by these drinks impair performance in tasks requiring both sustained attention and speed, and that the impairment correlates with the severity of the hangover, but not the colour of the drink.

The colours of alcoholic drinks are often due to congeners — compounds other than alcohol (ethanol) that form during the fermenting process. Bourbon contains 37 times more congeners than vodka, for example. Although the main cause of hangover symptoms is ethanol, congeners are thought to make matters worse.

Rohsenow *et al.* assessed this theory in a controlled study in which subjects drank vodka or bourbon until their blood alcohol concentration reached an intoxicating level. The intensity of the subjects' hangovers was quantified the next day using a symptom-based scale, and the subjects were given tests to measure their performance in tasks requiring speed and/or sustained concentration. Because alcohol affects the quality and duration of sleep, the authors also monitored these effects in their subjects.

Sure enough, bourbon caused worse hangovers than vodka, and all of the subjects experienced lighter, more disturbed sleep after alcohol consumption than they did when given a placebo (decaffeinated cola). But the effects of bourbon on sleep and on next-day performance in tests were no worse than those of vodka. What's more, although the amount of alcohol-induced sleep disturbance correlated with hangover severity, it was not responsible for the effects on performance in the tests.

Older subjects, or those dependent on alcohol, might have behaved differently from the young, healthy people who took part in the research. Nevertheless, as the authors point out, these findings have implications for people working in safety-sensitive jobs — as well as providing insight into how you feel the morning after the night before.
Andrew Mitchinson

---

stars strike each other, they are likely to merge into a single star, in an event characterized by relatively little mass loss[5]. This violent event, which could explain how blue stragglers came to be (Fig. 1a), may stir up the contents of the newly formed star, providing large amounts of hydrogen in the centre to be used as fuel in nuclear-fusion reactions. It is almost as if the star is reborn.

But collisions are unlikely to be the entire story, because blue stragglers have been seen in isolation in the dark-matter halo of our Galaxy, where stars are so sparse that close encounters between them will not occur sufficiently often to account for the number of blue stragglers seen there[6]. An alternative way to make a blue straggler is through mass transfer within a binary star system (Fig. 1b). These systems are very common; the Sun may be rather unusual in not having a stellar companion. If one of the stars in a binary star system runs out of hydrogen fuel in its core, it evolves from the main-sequence stage and its outer envelope expands — the star becomes a red giant. Because the star is now much larger, gas may flow from its envelope on to the companion star, provided the two stars are close enough. Even though the two stars never collide, the effect of the mass flow could be the same as a collision: the companion star becomes more massive and receives a fresh supply of hydrogen fuel to its core. After a phase of stable mass flow from one star to another, the system should contain a blue straggler orbiting around the remnant core of the donor star, in a wider binary system that has an orbital period of longer than one year. This seems to be the case for at least some of the blue stragglers seen in the halo of the Milky Way.

Using Hubble Space Telescope observations of the globular cluster M 30, Ferraro and colleagues[1] suggest that the blue stragglers in

M 30 have been made in these two distinct ways. They show that the blue stragglers in the cluster follow two separate sequences on the colour–magnitude diagram: one produced by collisions and the second produced by mass transfer within binaries. In both cases, a low-mass star is turned into a more-massive star that is about one-and-a-half times the mass of the Sun. If this happened relatively late in the cluster's life, then the more-massive star may not yet have had time to evolve to become a white dwarf, and so the star is seen today as a blue straggler.

Meanwhile, by analysing data obtained by the WIYN 3.5-metre telescope of the less-dense star cluster NGC 188, called an open star cluster because its stars are loosely bound by gravity, Mathieu and Geller[2] show how most of the cluster's blue stragglers are in binary systems with orbital periods of about 1,000 days.

This suggests that, in this cluster, most blue stragglers have been produced by mass flow between stars within a binary.

Taken together, these results[1,2] produce a consistent story: one in which blue stragglers are formed in both ways in star clusters, with collisions becoming more common with increasing cluster density, and binaries alone producing the blue-straggler population seen in the Galactic halo. ∎

Melvyn B. Davies is at the Lund Observatory, SE-221 00 Lund, Sweden.
e-mail: mbd@astro.lu.se

1. Ferraro, F. R. *et al. Nature* **462,** 1028–1031 (2009).
2. Mathieu, R. D. & Geller, A. M. *Nature* **462,** 1032-1035 (2009).
3. Sandage, A. R. *Astron. J.* **58,** 61–75 (1953).
4. Hills, J. G. & Day, C. A. *Astrophys. Lett.* **17,** 87–93 (1976).
5. Benz, W. & Hills, J. G. *Astrophys. J.* **323,** 614–628 (1987).
6. Carney, B. W., Latham, D. W., Laird, J. B., Grant, C. E. & Morse, J. A. *Astron. J.* **122,** 3419–3435 (2001).

STRUCTURAL BIOLOGY
Translocation chamber's secrets

Peter J. Christie

DNA transfer across membranes is a fundamental life process. The structure of part of a protein channel that performs this task offers insight into the mechanism of DNA passage through bacterial cell envelopes.

Many bacterial species transfer DNA to neighbouring cells in a process called conjugation. Over evolutionary time, conjugation has played a dominant part in shaping bacterial genomes. On a more immediate timescale, conjugation poses an enormous public-health problem as a mechanism underlying the rapid

dissemination of antibiotic-resistance genes and other virulence traits among pathogens. The question of how bacteria move DNA from one cell to the next has been investigated for more than 50 years, but only recently have structural details emerged. On page 1011 of this issue, a group led by Gabriel Waksman[1]

reports the first crystal structure of a large portion of a conjugation 'machine'. The researchers' findings represent a great leap in our understanding of how these transfer systems function at the molecular level.

Conjugation systems are a subfamily of type IV secretion (T4S) systems, whose range of functions includes DNA exchange with the environment and delivery of DNA and protein effectors to bacterial, fungal, plant and human target cells[2]. A few T4S systems in Gram-negative bacteria have been extensively studied, and one of these, the VirB/VirD4 T4S system from *Agrobacterium tumefaciens*, serves as a point of reference for this superfamily of translocation systems. In *A. tumefaciens*, 11 VirB subunits assemble together with a VirD4 subunit to form two structures: a gated channel for substrate transfer and a conjugative pilus (a hairlike appendage) for establishing contact with target cells[2].

During the past decade, Waksman and colleagues have solved the structures of several subunits of T4S systems[3–8], adding crucial details to an emerging picture of channel architecture. Earlier this year, these investigators presented a structure[9] of a portion of a conjugation channel, known as the core complex, from *Escherichia coli,* solved at a resolution of about 15 Å using cryo-electron microscopy (cryo-EM). The structure showed that 14 copies of the complex's TraN, TraO and TraF subunits — which are functionally equivalent to VirB7, VirB9 and VirB10 in *A. tumefaciens* — assemble as a large chamber, 185 Å wide and high, that spans the bacterial cell envelope (Fig. 1). The authors designated the portions of the chamber that are close to the inner and outer membranes of the envelope as inner (I) and outer (O) layers, respectively.

Waksman and colleagues[1] now provide the first high-resolution (2.6 Å) three-dimensional crystal structure of a portion of the core complex from *E. coli* (Fig. 1). The structure, which corresponds to the O-layer of the cryo-EM structure[9], is a tour de force in structural biology. It is the first X-ray structure of an outer-membrane pore that consists of more than one type of subunit; all other known outer-membrane pores or channels consist of variable numbers of a single subunit. It is also one of the largest membrane-protein complexes solved to date.

Surprisingly, the structure reveals that the outer-membrane pore, called the cap, is composed not of TraO (VirB9), as previously predicted[9], but of a domain of TraF (VirB10) known as the antennae projection[6]. This projection consists of two α-helices and an intervening loop; 14 copies of these two-helix bundles arrange to form the membrane-spanning pore. This is a strikingly unusual structure — almost all other bacterial outer-membrane proteins assemble as β-barrels in the membrane, rather than as α-helices.

Even more intriguingly, the new structure[1], taken together with earlier findings[10],

establishes that TraF (VirB10) spans the cell envelope. This is unprecedented among bacterial membrane proteins, and is especially fascinating because VirB10 is structurally dynamic[11]. Specifically, VirB10 senses conformational changes that accompany either the binding or hydrolysis of ATP (an energy source) by protein subunits situated at the conjugation channel's entrance. VirB10 responds to these changes by undergoing its own structural transition, which is required for the conjugation system to transfer a substrate. In the context of Waksman and colleagues' new structure[1], this ATP-sensing mechanism could serve to modulate gating of the α-helical pore to permit substrate passage. Indeed, the authors suggest that structural differences detected in their cryo-EM[9] and X-ray structures[1] might reflect different energy states of the complex.

So what is the large chamber in the core complex for? In *A. tumefaciens*, besides the VirB7–VirB9–VirB10 core complex, at least seven other subunits are required for substrate transfer[12]. These subunits might assemble within the core chamber, the dimensions of which are sufficiently large to house an intact translocation channel (Fig. 1a). Additionally

(or alternatively), the core chamber could serve as a staging arena for pilus biogenesis (Fig. 1b).

The idea that the core chamber acts as a structural scaffold for channel and/or pilus assembly is gratifyingly compatible with most biochemical data. But such a model needs to be reconciled with at least three experimental observations. First, it is known that deletion of at least half of the amino acids in the VirB10 antennae projections (which come together to form the outer-membrane pore) selectively abolishes pilus biogenesis without disrupting substrate transfer[10]. If substrates pass through the core chamber (Fig. 1a), how can they transit through the outer membrane without an intact pore?

Second, although most T4S-system substrates engage with their T4S systems at the channel entrance in the cytoplasm, some substrates — such as the pertussis toxin, a pentameric protein complex — first assemble in the periplasm (the space between the inner and outer membranes of the bacterial cell envelope) and then engage with the T4S system for transfer across the outer membrane[13]. But Waksman and colleagues' X-ray structure[1] reveals an extensive network of interactions
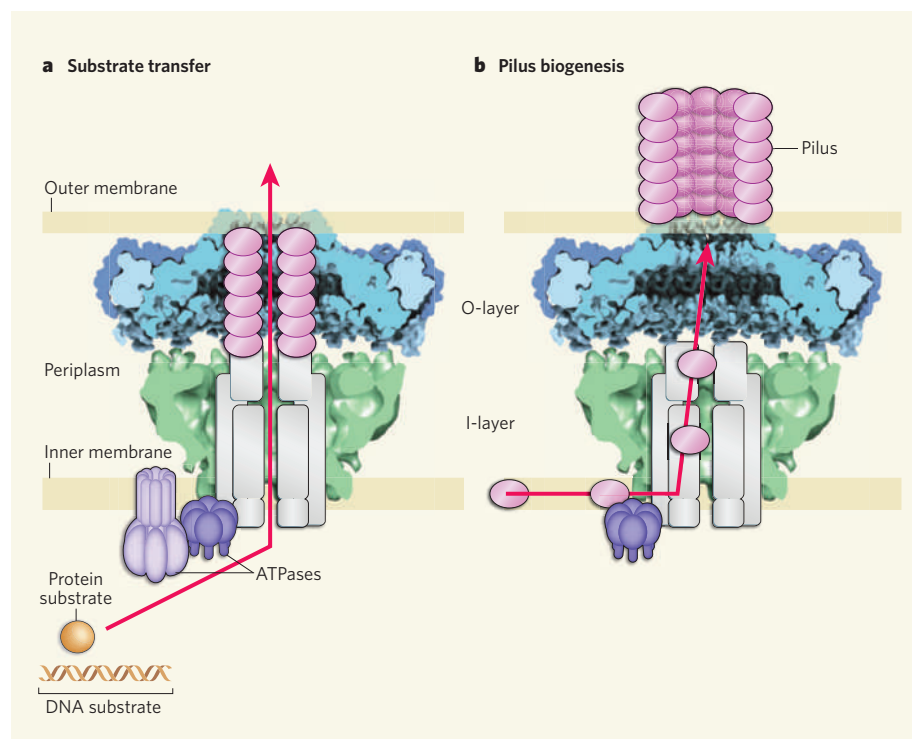

Figure 1 | Architecture and possible functions of a type IV secretion system. Bacteria transfer DNA and protein substrates to each other and to eukaryotic cells, such as those of plants and animals, using type IV secretion (T4S) systems. These systems span the periplasm region between the outer and inner membranes of bacterial cell envelopes. T4S systems contain a core complex that consists of two layers, the O-layer and the I-layer. The X-ray crystal structure of an O-layer (blue) reported by Waksman and colleagues[1] is shown here with a cryoelectron-microscope image[9] (green) of the same core complex. The remaining subunits of the system are depicted in schematic form. Waksman and colleagues propose two possible functions for the core complex. **a,** It might serve as a structural scaffold that houses the translocation channel through which substrates pass. Several channel subunits, including pilin proteins (magenta ovals), form a conduit encased within the core chamber, and ATPase subunits bind the substrates ready for translocation. **b,** The core complex might also act as a scaffold and assembly platform for the conjugative pilus — the hair-like appendage of the T4S system that is used to make contact with other bacteria (or eukaryotic cells) prior to substrate transfer. Pilin proteins in the inner membrane would be passed through the core chamber to the bacterial cell surface to build the pilus.

among the core subunits of T4S systems, and it is not evident how multisubunit substrates could gain access to the core's chamber for translocation.

Finally, the diameter of the α-helical pore in the new crystal structure[1] is only 32 Å, a size that could accommodate DNA and unfolded protein substrates, but not a multisubunit protein substrate or, for that matter, the conjugative pilus (80–120 Å in diameter)[14] without inducing gross structural rearrangements.

These and other issues await further study, but in the meantime Waksman and colleagues are to be credited for a quantum leap in our understanding of macromolecular translocation in bacteria. On the structural front, the next task is to resolve even larger subassemblies of T4S systems and, ultimately, the entire translocation machine. Defining how the core complex physically associates with the conjugative pilus will also be a formidable technical challenge, but this is an essential goal if we are to understand the role, if any, of the pilus in substrate transfer. Equally necessary are functional studies to establish that the structures solved *in vitro* exist and have biological activity *in vivo*. Besides stimulating these and other avenues of research, Waksman and colleagues' findings[1] move us a step closer to developing drugs that target T4S systems, for controlling the spread of antibiotic resistance and mitigating the proliferation of medically important pathogens. ■

Peter J. Christie is in the Department of Microbiology and Molecular Genetics, University of Texas Medical School at Houston, Houston, Texas 77030, USA.
e-mail: peter.j.christie@uth.tmc.edu

1. Chandran, V. *et al. Nature* **462,** 1011–1015 (2009).
2. Alvarez-Martinez, C. E. & Christie, P. J. *Microbiol. Mol. Biol. Rev.* **73,** 775–808 (2009).
3. Yeo, H.-J., Savvides, S. N., Herr, A. B., Lanka, E. & Waksman, G. *Mol. Cell* **6,** 1461–1472 (2000).
4. Savvides, S. N. *et al. EMBO J.* **22,** 1969–1980 (2003).
5. Yeo, H.-J., Yuan, Q., Beck, M. R., Baron, C. & Waksman, G. *Proc. Natl Acad. Sci. USA* **100,** 15947–15952 (2003).
6. Terradot, L. *et al. Proc. Natl Acad. Sci. USA* **102,** 4596–4601 (2005).
7. Hare, S., Bayliss, R., Baron, C. & Waksman, G. *J. Mol. Biol.* **360,** 56–66 (2006).
8. Bayliss, R. *et al. Proc. Natl Acad. Sci. USA* **104,** 1673–1678 (2007).
9. Fronzes, R. *et al. Science* **323,** 266–268 (2009).
10. Jakubowski, S. J. *et al. Mol. Microbiol.* **71,** 779–794 (2009).
11. Cascales, E. & Christie, P. J. *Proc. Natl Acad. Sci. USA* **101,** 17228–17233 (2004).
12. Cascales, E. & Christie, P. J. *Science* **304,** 1170–1173 (2004).
13. Burns, D. L. *Curr. Opin. Microbiol.* **6,** 29–34 (2003).
14. Wang, Y. A., Yu, X., Silverman, P. M., Harris, R. L. & Egelman, E. H. *J. Mol. Biol.* **385,** 22–29 (2009).

NANOTECHNOLOGY

Molecular transistors scrutinized

James Kushmerick

Transistors have been made from single molecules, where the flow of electrons is controlled by modulating the energy of the molecular orbitals. Insight from such systems could aid the development of future electronic devices.

Transistors, the fundamental elements of integrated circuits, control the flow of current between two electrodes (the source and drain electrodes) by modifying the voltage applied at a third electrode (the gate electrode). As manufacturers compete to produce ever smaller devices, one logical limit to circuit miniaturization is transistors whose channels are defined by a single molecule. The construction and characterization of such a device has been a long-standing goal of nanoelectronics.

In fact, solid-state molecular transistors have already been made, and are based on two mechanisms: Coulomb blockade, in which the flow of electrons is controlled by the sequential charging of a molecule; and the Kondo effect, in which conducting electrons interact with local spin (intrinsic angular momentum) in a molecular junction[1–4]. A third approach has been predicted[5], based purely on electrostatic modulation of the molecular-orbital energy of a single molecule. On page 1039 of this issue, Song *et al.*[6] describe the first practical realization of this approach, and characterize the

resulting devices in unprecedented detail.

Despite various reports[1–4,7] of single molecules being electrically connected to two electrodes, achieving this feat is a daunting task. Song *et al.*[6] fabricated their devices by coating gold wires with the molecules of interest, then breaking the wires using a technique called electromigration and thus producing nanometre-scale gaps in the wires (Fig. 1). On occasion, the molecules coating the wire fortuitously become trapped in the gaps. This results in systems of source/drain electrodes — two broken ends of the wire — spanned by the molecules, forming a junction through which electrons can 'tunnel'. One of the benefits of this fabrication method is that the junction can be formed directly over an oxidized-aluminium gate electrode, thus providing the necessary three-terminal geometry of a transistor.

Where Song and colleagues' study excels is in the detailed examination of their molecular junctions. The field of molecular electronics has long been plagued by concerns that the observed current–voltage characteristics

of reported devices are caused by impurities or defects in the systems, rather than by the molecular species under study. Song et al.[6] have avoided such uncertainties by thoroughly characterizing the charge-transport properties of their devices using a combination of spectroscopy techniques in situ. In this way, they provide unprecedented insight into the underlying physics of charge transport in their molecular transistors.

The authors used inelastic electron tunnelling (IET) spectroscopy to measure the interactions between the tunnelling electrons and the vibrational modes of the molecules in their devices. This technique provides definitive proof that the measured currents actually pass through the molecules in single-molecule transistors, and yields some information about the pathways taken by electrons as they cross the junctions[8]. The authors tested two types of transistor, each with a different molecule in the junction — either an alkane dithiol (which contains two SH groups connected by a saturated hydrocarbon chain) or an aromatic dithiol (which contains two SH groups connected by a benzene ring). Because each dithiol has its own vibrational 'fingerprint', the IET spectra of the devices provide unambiguous evidence of the molecules in the junctions.

The second technique used by Song et al. was transition-voltage spectroscopy. Electrons crossing a molecular junction do so using one of two tunnelling mechanisms that depend on the magnitude of the source–drain voltage; the transition voltage (V_{trans}) is the voltage at which tunnelling switches from one mechanism to the other. It has previously been shown[9] that V_{trans} is proportional to the difference in energy between the gating orbital of the molecular junction (the orbital that modulates electron tunnelling) and the Fermi levels of the source and drain electrodes, where the Fermi level is the highest possible energy for a conducting electron in an electrode. By measuring V_{trans} using transition-voltage spectroscopy at different applied gate voltages, Song et al. demonstrated that a linear relationship exists between gate voltage and molecular-orbital energy in their devices, as expected for single-molecule transistors.

The nature of the molecular orbital that couples to the tunnelling electrons (that is, whether or not the orbital is occupied by electrons from the molecule in the junction) can be determined from the change in conductance of the transistor with respect to the applied gate voltage. Song et al. found that both of their transistor types become more conducting when a negative gate voltage is applied. Because negative gate voltages lower the energy difference between the highest occupied molecular orbital (HOMO) of the molecular junction and the electrode's Fermi level, this indicates that, in their device, tunnelling electrons couple to the HOMOs of the molecules.

The authors found further evidence that the current through their device was gated by the HOMO energy of the molecular junction by
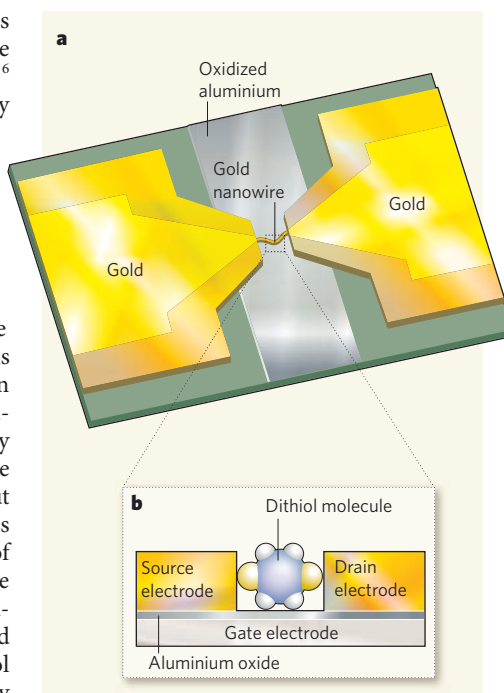
Figure 1 | Single-molecule transistors. Song et al.[6] have made and characterized single-molecule transistors in which current flow is controlled by electrostatically modulating the energy of the molecular orbitals of a single molecule. **a**, Each device consists of a fractured gold nanowire overlaid on a strip of oxidized aluminium. **b**, Side-on, close-up view of a device. The broken ends of the nanowire form the source and drain electrodes of the transistor, and the oxidized aluminium forms the gate electrode. Aluminium oxide on the surface of the gate electrode provides a necessary layer of insulating material known as the gate dielectric. A single molecule (here, an aromatic dithiol) connects the source and drain electrodes. The electric field created by the gate electrode modulates the energies of molecular orbitals in the dithiol, which in turn control the amount of current that flows through the source and drain electrodes. The components of the device are not drawn to scale.

examining the dependence of the IET spectra on the applied gate voltage. The IET spectra of the transistors that incorporate alkane dithiols were essentially unaffected by the gate voltage. This indicates that electron tunnelling through the device is always 'non-resonant', that is, there is a large energy difference between the dithiol's HOMO and the electrode's Fermi level.

Conversely, Song et al. observed that the applied gate voltage strongly modulates the IET spectra of transistors that incorporate an aromatic dithiol. Specifically, when a negative gate voltage is applied (which brings the energy of the molecular junction's HOMO closer to that of the electrode's Fermi level), the signal intensities of the spectra increase greatly and the shapes of the vibrational peaks change. The change in peak shape is a clear indication of increased coupling between the tunnelling charge carrier and the molecular vibrations, owing to a resonance between the HOMO and the Fermi

level[10]. The authors have thus provided the first experimental demonstration that resonant and non-resonant vibrational coupling can be tuned in single-molecule transistors.

One of the most surprising features of Song and co-workers' study[6] is the strong effect of the gate voltage on the molecular-orbital energy of their device. For both dithiols studied, the molecular orbitals shifted in energy by 0.25 electronvolts when 1 volt was applied to the gate electrode, a remarkably strong coupling. It is unclear why the gate coupling should be so strong, but the most likely explanation is that the molecules are extremely close to, or in intimate contact with, the gate dielectric (the oxidized aluminium of the gate electrode). The need for such precise alignment may in part explain why so few of the devices prepared by the authors functioned properly as transistors — only 35 out of 418 were found to have the desired current–voltage characteristics.

Through their multi-spectroscopy approach, Song et. al.[6] have provided the first conclusive evidence that a solid-state molecular transistor can function through the relative alignment of its orbital energies with the electrode's Fermi level, and that this alignment can be efficiently tuned by the applied gate voltage. Their work sets a benchmark for the validation of future studies of charge transport in molecular systems. But much work remains to be done before molecular electronic devices can effectively compete with their larger silicon-based brethren. For example, a fabrication method that provides high yields of densely packed single-molecule devices has yet to be developed. In the meantime, Song and colleagues' work provides an excellent foundation for further development of well-characterized molecular devices. ■

James Kushmerick is at the National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA.
e-mail: james.kushmerick@nist.gov

1. Park, J. et al. Nature **417,** 722–725 (2002).
2. Liang, W., Shores, M. P., Bockrath, M., Long, J. R. & Park, H. Nature **417,** 725–729 (2002).
3. Kubatkin, S. et al. Nature **425,** 698–701 (2003).
4. Yu, L. H. & Natelson, D. Nano Lett. **4,** 79–83 (2004).
5. Ghosh, A. W., Rakshit, T. & Datta, S. Nano Lett. **4,** 565–568 (2004).
6. Song, H. et al. Nature **462,** 1039–1043 (2009).
7. Xu, B. & Tao, N. J. Science **301,** 1221–1223 (2003).
8. Troisi, A. et al. Proc. Natl Acad. Sci. USA **104,** 14255–14259 (2007).
9. Beebe, J. M., Kim, B.-S., Gadzuk, J. W., Frisbie, C. D. & Kushmerick, J. G. Phys. Rev. Lett. **97,** 026801 (2006).
10. Galperin, M., Ratner, M. A. & Nitzan, A. J. Chem. Phys. **121,** 11965–11979 (2004).

Word of the Year
The News & Views WotY for 2009 is 'critical'. But this is not an accolade. This devilish little word has an honourable place in the scientific lexicon when bound together with such terms as 'mass', 'angle' and 'temperature'. But in an evolutionary battle to convey the sense of 'having a decisive importance', it has all but killed off such alternatives as 'crucial', 'vital', 'essential' and 'necessary'. We can but hope that its spread is halted in 2010.

OBITUARY

Vitaly Ginzburg (1916–2009)

Nobel-prizewinning success in physics achieved in the Soviet system.

Vitaly Lazarevich Ginzburg, who died on 8 November, played a leading part in many aspects of theoretical physics during the Soviet era and after the dismantling of the Soviet Union in 1991. His research contributions were vast and of the highest order, culminating in the award of the Nobel Prize in Physics in 2003, jointly with Alexei Abrikosov and Anthony Leggett, for pioneering studies in superconductivity and superfluidity.

Ginzburg was born in 1916 into a Jewish family in Moscow just before the Russian Revolution. His formal school education began only at the age of 11. In 1931, Evgeni Bakhmeťev, a professor at Moscow's Technical University, helped him to get a job as a laboratory assistant in the university's X-ray laboratory. This experience whetted his appetite for physics and he entered Moscow State University in 1933 to study this discipline. Strongly attracted to theoretical physics, he was unsure of his mathematical ability and decided to work in optics under the supervision of Grigory Landsberg. He went on to take his PhD in 1940 and, having transferred to the Lebedev Physical Institute of the Soviet Academy of Sciences (FIAN) in Moscow, completed his science doctorate there in 1942. He was to remain a member of the FIAN for the rest of his life.

The scope of Ginzburg's research can be appreciated from his own attempt at a scientific autobiography, in which he listed, roughly chronologically, his range of interests in theoretical physics: classical and quantum electrodynamics, Cherenkov and transition radiation, the propagation of electromagnetic waves in plasma, radio astronomy and synchrotron radiation, cosmic-ray and γ-ray astrophysics, the scattering of light in crystals, the theory of ferroelectrics, and superfluidity and superconductivity.

In all of these areas, he wrote prolifically and made original contributions. For example, his work on transition radiation, a phenomenon that occurs when high-speed charged particles cross two media of different electric permittivity, followed on from his deep interest in electrodynamics, and his comprehensive treatment of the topic was truly pioneering. His studies of synchrotron radiation were highly influential in establishing that this process is the dominant non-thermal radiation mechanism in high-energy astrophysical phenomena in radio astronomy. These diverse interests were reflected in a series of influential books.

The pinnacle of his scientific achievement was his groundbreaking research with Lev

A. ZEMLIANICHENKO/AP

Landau on the theory of superconductivity, published in 1950. This work built on Landau's theory of second-order phase transitions. Ginzburg had already applied Landau theory to ferroelectric phenomena. The crucial advance in the Ginzburg–Landau theory was the concept that, in the transition from the normal to the superconducting state, the phenomenon of symmetry breaking in a metal, a characteristic of Landau theory, was associated with the wavefunction of the metal's superconducting electrons, a non-gauge invariant process. This new paradigm was to have profound implications for many aspects of quantum physics beyond superconductivity, including the Higgs phenomenon, which gives particles mass.

His broad interests in theoretical physics were reflected in the famous Ginzburg seminars, which were held each week at the FIAN. Ginzburg stated that the topics for discussion should include all theoretical physics, except particle physics. His leadership at these seminars was impressive: he regularly interrupted the speaker to summarize what had just been said so that all listeners could follow the argument. These were celebrated weekly events, with most of Moscow's physicists making an effort to attend.

Ginzburg lived through a turbulent era. The Soviet Union entered the Second World War in 1941, and the Soviet Academy of Sciences was evacuated to Kazan, where Ginzburg worked for the next two years. In 1937, he had married his fellow student Olga Zamsha, but they divorced in 1946. In the same year he married Nina Ermakova, who had been arrested in 1944 on a trumped-up charge of plotting to kill Stalin. She was given a lenient sentence and released under an amnesty in 1945, but was not allowed to return to Moscow.

In 1945, Ginzburg was invited to become a visiting professor at the newly established radiophysical department at Gorky University in what is now Yekaterinburg, and he subsequently became chair of a group studying the propagation and radiation of radio waves. Living mainly in Moscow, for seven years Ginzburg made annual applications for his wife to be allowed to return there, but these were refused until after Stalin's death in 1953.

In 1947, he was personally attacked in an article in the *Literaturnaya Gazeta*, which blamed him for non-patriotic citations in his papers and for 'idealism'. Despite this attack, Igor Tamm, in need of physicists of the highest quality, arranged that he join the Soviet nuclear-weapons programme. The leaders of this project — Yulii Khariton, Igor Kurchatov and Yakov Zeldovich — assembled a brilliant team of physicists and mathematicians, including Andrei Sakharov, Israil Gel'fand, Alexander Kompaneets, Landau and Ginzburg, to develop nuclear weapons in response to the United States' development of the atomic and hydrogen bombs. Ginzburg's major contribution was to propose the use of lithium-6 as the fuel for the Soviet hydrogen bomb, a quite different process from that adopted in the United States. But he did not remain long in the nuclear programme.

After Stalin's death, Ginzburg was elected a corresponding member of the Soviet Academy of Sciences and his wife returned to Moscow. He became a full member of the academy in 1966 and, on the death of Tamm in 1971, became head of the theoretical physics department at the FIAN. During this period, Sakharov had become politically active and was classed as a dissident, being exiled to Gorky in 1980. The FIAN provided a scientific home for him, but, as a member of the theoretical physics department, this placed significant constraints on Ginzburg, who was not allowed to travel abroad for many years. Matters changed significantly with the period of perestroika that started in 1985, with Ginzburg being appointed a member of the Congress of People's Deputies from 1989 to 1991, when the body was dissolved.

Ginzburg was a strong personality, with deeply held humanitarian views that he maintained throughout the years of Soviet rule. He kept an open mind on issues in theoretical physics, but based his opinions on a strongly developed intuition for the underlying principles. He will be remembered with gratitude by all who experienced his kindness, and as an inspirational figure who carried out world-leading research against a background of significant political oppression.

Malcolm Longair
Malcolm Longair is at the Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK.
e-mail: msl1000@cam.ac.uk

nature

REVIEWS

The first decade of science with Chandra and XMM-Newton

Maria Santos-Lleo[1], Norbert Schartel[1], Harvey Tananbaum[2], Wallace Tucker[2] & Martin C. Weisskopf[3]

NASA's Chandra X-ray Observatory and the ESA's X-ray Multi-Mirror Mission (XMM-Newton) made their first observations ten years ago. The complementary capabilities of these observatories allow us to make high-resolution images and precisely measure the energy of cosmic X-rays. Less than 50 years after the first detection of an extrasolar X-ray source, these observatories have achieved an increase in sensitivity comparable to going from naked-eye observations to the most powerful optical telescopes over the past 400 years. We highlight some of the many discoveries made by Chandra and XMM-Newton that have transformed twenty-first century astronomy.

Typically, cosmic X-rays are produced in extreme conditions—from intense gravitational and magnetic fields around neutron stars and black holes, to intergalactic shocks in clusters of galaxies. Chandra[1] and XMM-Newton[2] have probed the space-time geometry around black holes, unveiled the importance of accreting supermassive black holes in the evolution of the most massive galaxies, demonstrated in a unique manner that dark matter exists, and confirmed the existence of dark energy. They have also tracked the production and dispersal of heavy elements by supernovae and measured the magnitude and rate of flaring of young Sun-like stars. Table 1 gives a subjective and by-no-means complete list of significant discoveries made using these observatories.

With an order-of-magnitude or more improvement in spectral and spatial resolution and sensitivity, Chandra and XMM-Newton have shed light on known problems, as well as opened new areas of research. These observatories have clarified the nature of X-ray radiation from comets[3], collected a wealth of data on the nature of X-ray emission from stars of all ages[4,5], and used spectra and images of supernova shock waves to confirm the basic gas dynamical model[6,7] of these objects. They have resolved into discrete point sources the diffuse emission from the plane of the Galaxy[8], as well as the diffuse extragalactic X-ray background[9]. They have discovered hundreds of supermassive black holes at the centres of galaxies and for many of those obtained high-resolution spectra that have

Table 1 | A sample of discoveries made using the Chandra and XMM-Newton observatories

Topic	Discovery
Comets	Established charge-exchange as mechanism for X-ray emission.
Individual stars	Measured densities, temperatures and composition of hot plasmas, testing models for stellar evolution, X-ray emission from stellar coronae, and stellar winds.
Star formation and star-forming regions	Discovered X-ray emission from gas accreting onto stellar surfaces and influenced by magnetic fields; detected giant flares from young stars, with implications for planet formation.
Supernovae	Established that Kepler's supernova was a thermonuclear event.
Supernova remnants (SNRs)	Discovered a central compact object in the Cas A SNR and traced the distribution of elements indicating turbulent mixing along with an aspherical explosion. Imaged forward and reverse shock waves in several SNR, with implications for the acceleration of cosmic rays.
Pulsar wind nebulae	Resolved jets and rings of relativistic particles produced by young neutron stars.
Black hole accretion processes	Provided evidence for rotation of space-time around black holes; measured the efficiency of the accretion process; and detected jets and winds produced by black holes.
Galactic Centre	Measured the flaring of central black hole and resolved the galactic ridge emission into individual sources.
Starburst galaxies	Discovered evidence for enrichment of the interstellar medium and the intergalactic medium by starbursts.
Supermassive black holes and active galactic nuclei (AGNs)	Resolved the X-ray background radiation into discrete sources, mostly supermassive black holes; traced the history of supermassive black hole growth over cosmic timescales.
Active galactic nuclei feedback in galaxies and clusters of galaxies	Discovered evidence for heating of hot gas in galaxies and clusters by outbursts produced by supermassive black holes, supporting the concept that supermassive black holes can regulate the growth of galaxies.
Dark matter	Determined the amount of dark matter in galaxy clusters and, by extension, the Universe; observed the separation of dark matter from normal matter in the Bullet Cluster, demonstrating that alternative theories of gravity are very unlikely to explain the evidence for dark matter.
Dark energy	Observed galaxy clusters to generate two independent measurements of the accelerated expansion of the Universe.

[1]XMM-Newton Science Operations Centre, European Space Agency, Villanueva de la Cañada, 28691 Madrid, Spain. [2]Chandra X-ray Center, Smithsonian Astrophysical Observatory, Cambridge, Massachusetts 02138, USA. [3]Space Science Office, NASA Marshall Space Flight Center, Huntsville, Alabama 35812, USA.
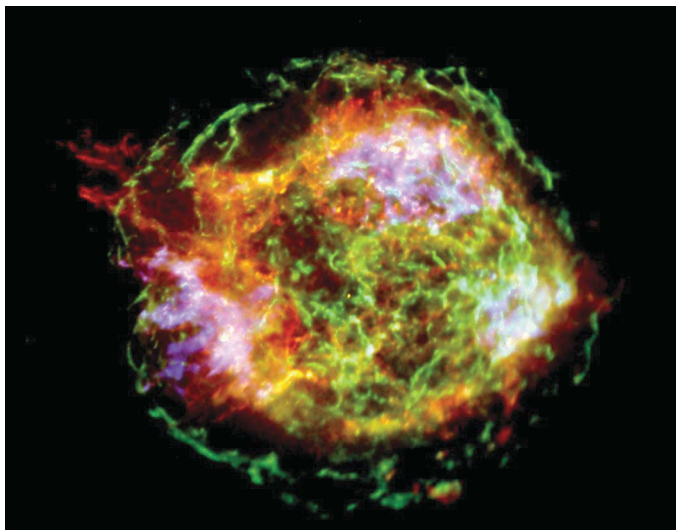
Figure 1 | Chandra image of the Cassiopeia A supernova remnant. Red, green and blue denote lower, medium and higher X-ray energies, respectively. Chandra's sub-arcsecond imaging clearly delineates the shock wave (bright green outer ring) generated by the supernova. A large jet-like structure (mostly red) protrudes beyond the shock wave towards the upper left. Reproduced with permission from ref. 89.

provided insight into the accretion process that powers the activity[10,11].

Within the Galaxy, surprises from Chandra and XMM-Newton included spectacular images of the Cassiopeia A (Cas A) supernova remnant (Fig. 1), of the Crab nebula (revealing rings and jets around the rapidly rotating neutron star[12]), and of numerous other pulsar wind nebulae[13]. The observatories found diverse X-ray emission mechanisms operating in normal stars[4]. Beyond the Galaxy, X-ray

Figure 2 | Composite of XMM-Newton optical-ultraviolet and X-ray images of the starburst galaxy Messier 82. The optical-ultraviolet image—superimposed on the X-ray image (blue)—shows emission from the plane of this galaxy, obscuring dust lanes, and bright knots indicative of intense star formation. The X-ray image shows plumes of hot gas bursting from the galactic disk, in an extensive outflow driven by a burst of star formation. The X-ray emission is a combination of a continuum from many point sources in M82's central region and from collisionally ionized plasma whose temperature and abundance change along the outflow, plus emission lines due to charge-exchange reactions involving neutral metals (Mg and Si) in dust grains[90]. This composite image illustrates how simultaneous observations with XMM-Newton's on-board optical-ultraviolet camera enable extended wavelength coverage. Reproduced with permission from P. Rodriguez and the European Space Agency (ESA).

and radio data demonstrated convincingly that supermassive black holes and intense bursts of star formation can affect their environment (Fig. 2) on scales of hundreds of thousands of light years[14]. X-ray and optical data revealed the separation of dark and normal matter in a cosmic collision[15], and galaxy cluster surveys provided independent evidence for the accelerated expansion of the Universe[16,17].

Solar System objects

The interaction of energetic particles and radiation from the Sun with the atmospheres of planets, their satellites, and comets produces X-ray radiation through scattering, fluorescence, charge exchange, or the stimulation of auroral activity. Chandra and XMM-Newton have observed—and in many cases discovered—X-rays from Venus[18]; the Earth[19] and its moon[20]; Mars[21]; Jupiter[22,23], its aurorae[3] and disk[24], some of its moons[3,23] and the Io plasma torus[23]; Saturn[25–27] and its rings[28]; and numerous comets[29,30]. For detailed overviews of these observations, see references 3, 22 and 23.

An important mechanism for producing X-rays from Solar System objects is charge exchange, which occurs when a highly ionized atom in the solar wind collides with a neutral atom (gas or solid) and captures an electron, usually in an excited state. As the ion relaxes, it radiates an X-ray characteristic of the wind ion. Lines produced by charge exchange with solar wind ions such as C V, C VI, O VII, O VIII and Ne IX have all been detected with Chandra and XMM-Newton (for example, see Fig. 3).

The X-ray spectrum of a comet measures its out-gassing rate[29,30] and probes the solar wind *in situ*—in essence, a laboratory for highly ionized particles in low-density plasmas. Before XMM-Newton and Chandra, the origin of cometary X-rays was debated. Now, observational and theoretical work has demonstrated that charge-exchange collisions of highly charged solar wind ions with cometary neutral species are the best explanation for the emission[3].

Chandra and XMM-Newton have also discovered X-ray signatures of charge exchange in the exospheres of Venus and Mars, thus enabling measurement of planetary out-gassing rates, remotely and on a global scale. In addition, these observatories probe the planetary atmospheres through fluorescence and scattering of solar X-rays by constituents of the atmosphere (Fig. 3).

Planets with a magnetic field, such as Earth, Jupiter and Saturn, generate X-rays by auroral activity. Chandra discovered[23] time-variable X-ray flux from Jupiter that originates, not from the ultraviolet auroral
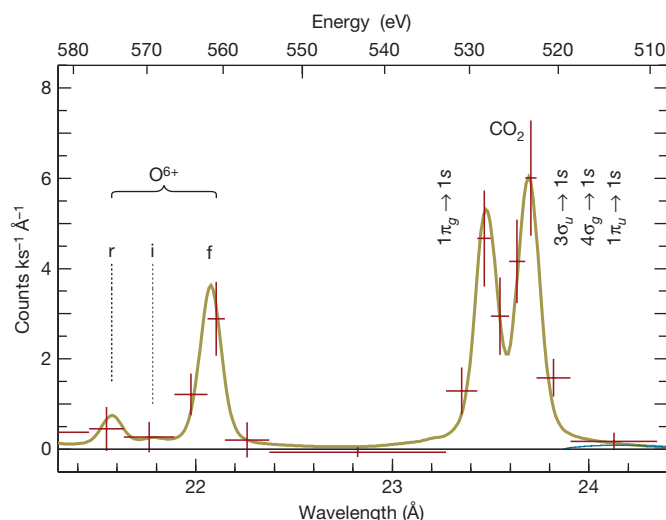

Figure 3 | Portion of an XMM-Newton reflection-grating spectrum of the disk of Mars. The spectrum shows prominent oxygen lines from two processes: charge exchange of solar-wind ionic oxygen (left) and fluorescence by solar X-rays of bound oxygen in the predominately CO_2 Martian atmosphere (right). Uncertainties are 1 standard deviation statistical errors and the solid line is a theoretical model. The symbols r, i and f refer to the resonance, intercombination and forbidden lines, respectively. Reproduced with permission from K. Dennerl (personal communication).

zone as had been expected, but from higher latitudes near Jupiter's pole—regions that map to the outer boundary of Jupiter's magnetosphere. Charged particles precipitating into the Jovian atmosphere must therefore originate from much farther than the Io plasma torus. In addition, XMM-Newton observations[22] detected a component of the emission above 2 keV that is probably due to electron bremsstrahlung. Altogether these discoveries suggest a complicated current system in Jupiter's polar magnetosphere, requiring an important revision of the models.

Chandra and XMM-Newton also detected charge-exchange X-rays from interaction of the solar wind with Earth's exosphere (geocorona), with material in and outside the magnetosheath[31], and possibly with the heliosphere[32]. The data suggest that at X-ray energies near 0.75 keV the total local soft X-ray intensity is from the heliosphere, whereas near 0.25 keV, the heliospheric component is significant, but does not explain all the observations. A mix of solar wind charge-exchange emission and a warm ($\sim$0.6 MK), rather than hot, local bubble in the interstellar medium seems to be needed[33]. This discovery affects the search for X-ray emission from the warm, hot, intergalactic medium (0.1 MK to 10 MK) plasma that has been proposed to contain over half the baryons at the current epoch. Ignoring the contribution of solar wind charge-exchange emission to the background could lead to the wrong interpretation of observed emission as due to the warm, hot, intergalactic medium[31].

Individual stars and star-forming regions

Although the X-ray luminosity of stellar coronae is but a small fraction of the stellar bolometric luminosity, X-ray observations have proven to be good indicators of magnetic activity and the effectiveness of stellar magnetic dynamos. High-resolution X-ray spectroscopy with the objective gratings on XMM-Newton and Chandra have enabled measurement of many spectral lines—including resolved multiplets—and, in some cases, line profiles and line shifts. These measurements provide critical diagnostics of densities, temperatures and composition of hot coronal plasmas[4]. This capability has now enabled tests of models for X-ray emission from stellar coronae and, in the process, raised some intriguing questions. X-ray spectra of the coronae of cool stars have revealed an unexpected correlation between coronal abundance ratios and stellar magnetic activity and evolution. For example, the Ne/Fe ratio increases by an order of magnitude from the least active to most active coronae and even changes during a flare[4]. This effect is opposite to that seen in the solar corona, and is not yet completely understood.

Before XMM-Newton and Chandra, the mechanism for X-ray emission from massive, hot stars was generally thought to be due to shock waves colliding in the stellar wind, far above the stellar surface. The new data, although confirming some aspects of that model, show that adjustments are needed; for example, estimates for the mass-loss rate of massive stars may need to be revised downwards, or large-scale shocks from magnetically confined hot plasma may have a more important role than previously thought[4].

Two projects have been dedicated to study well-known, nearby, star-formation regions with many young stars: the XMM-Newton Extended Survey of the Taurus Molecular Cloud[34] (XEST) and the Chandra Orion Ultradeep Project[35,36] (COUP).

The Taurus molecular cloud is a test-bed for low-mass ($\sim$1 solar mass) star formation in dark clouds without luminous O-type stars. In this environment, stars form in isolation or in very small groups. The Taurus molecular cloud contains many T Tauri stars—very young stars that are still undergoing gravitational contraction and represent an intermediate stage between a protostar and a middle-aged star like our Sun.

Grating spectra from XEST observations have revealed that classical T Tauri stars accreting from circumstellar disks show excess flux in emission lines formed at low temperature—from, for example, C, N, O (Fig. 4)[4,37]. This 'soft X-ray excess' is thought to arise from magnetically collimated accretion columns at the stellar surface. Furthermore,
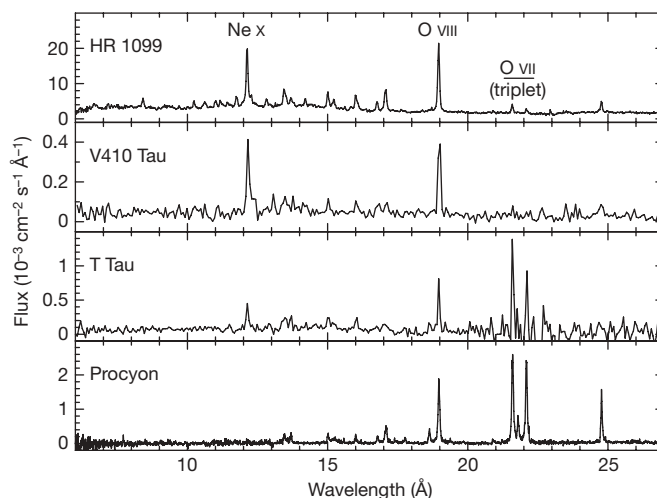

Figure 4 | XMM-Newton X-ray spectra of different types of stars. From top to bottom: the active binary HR 1099; the weak-line T Tauri V410 Tau; the classical T Tauri T Tau after absorption correction; and the inactive main sequence star Procyon. The O VII line, forming at 1–3 MK, is the strongest line in the X-ray spectrum of T Tau, despite the presence of large amounts of hot, coronal gas at 10–20 MK. Reproduced with permission from ref. 37.

the soft X-ray excess may be an important source of ionizing radiation for protoplanetary disks.

In contrast with the Taurus molecular cloud, the Orion nebula cluster is the prototype for 'clustered' star formation, where hundreds to thousands of low-mass stars form around massive O-type stars. The COUP image (Fig. 5) reveals X-ray emission from brown dwarfs, protostars with a wide range of ages, and stars ranging from solar-type to massive O-type. X-ray flares are seen at all stages, from protostars to older systems without circumstellar disks.

The COUP survey found that X-ray flares are ubiquitous in pre-main-sequence and young stars. Flares result from reconnection of multipolar surface magnetic fields (as on the Sun), but field lines reaching the protoplanetary disk may be involved[5,38]. Pre-main-sequence stars produce powerful flares that are about 100 times more
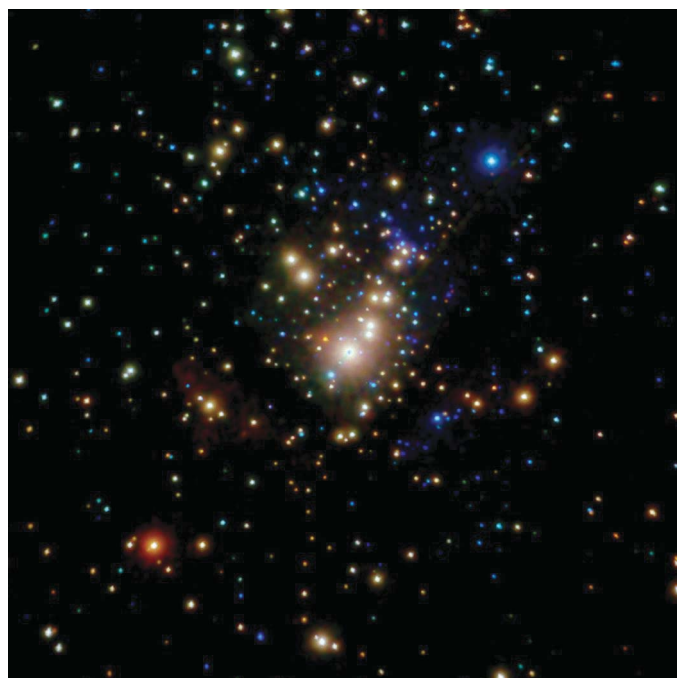

Figure 5 | Deep Chandra image of the Orion nebula cluster. This image shows more than a thousand young stars. Colours denote X-ray energy band: low (0.2–1.0 keV) in red, medium (1.0–2.0 keV) in green, and high (2.0–8.0 keV) in blue. Reproduced with permission from ref. 35.

intense and 100 times more frequent than solar flares[39]. X-rays from young-star flares may also dominate the ionization of a circumstellar protoplanetary disk, thereby altering its structure, dynamics and chemistry[40]. These processes may affect the formation of planets and their habitability.

The southwest side of Orion, a nearly circular region of much lower brightness than the northeast, is dominated by the massive stars. XMM-Newton observations revealed that this region is filled by hot (~2 MK) plasma[41], produced by interaction of the hypersonic wind from the most luminous Orion nebula cluster star (θ^1 Ori C) with the surrounding cold molecular cloud[42]. Detection of this plasma and similar structures in this and other regions with massive young stars reveals the large-scale effect of O-type star winds. Hot plasma from shocked winds leaking from the Orion nebula cluster feed a network of channels and bubbles of hot plasma in the Orion molecular cloud.

Supernova remnants and massive stellar black holes

The final stage for a massive star is a supernova caused by collapse of the star's central core to form a neutron star (and then a black hole if the star is sufficiently massive). The enormous gravitational energy released during the collapse violently ejects most of the star at speeds of thousands of kilometres per second. The expanding supernova remnant (SNR) disperses heavy elements and generates shocks that heat ejected and ambient material, as well as accelerate particles to relativistic energies.

Chandra, Hubble Space Telescope and Spitzer observations of SN 1987A at different epochs trace the detailed evolution of this very young SNR (Fig. 6). X-rays are produced by shocks encountering dense fingers that protrude inward from a circumstellar ring of material. X-ray spectral line ratios measure the temperature and chemical abundance of shocked material[43], and line profiles give kinematic velocities of the gas[44]. Modelling of the Chandra images over a 10-year period shows clear evidence for the deceleration of the expansion as the remnant encounters circumstellar material[45].

Chandra and XMM-Newton have also enabled detailed studies of the temperature, chemical composition and structure of SNRs. For example, the Chandra image of the Cas-A SNR (Fig. 1) shows iron-rich ejecta outside silicon-rich ejecta, thus indicating that turbulent mixing and an aspherical explosion turned much of the original star 'inside out'[46]. Observations of Doppler-shifted emission lines for Cas A and other SNRs are providing three-dimensional information on the distribution and velocity of the supernova ejecta which will help to constrain models for the explosion[47]. The image of Cas A also revealed a point source that is probably a neutron star with a 10-cm-thick atmosphere of carbon[48].

In contrast to core-collapse supernovae of massive stars, a thermonuclear (type Ia) supernova occurs if a white-dwarf star becomes unstable by accreting too much matter or merging with another white dwarf. X-ray observations of the abundance and distribution of iron and oxygen in the Kepler SNR show that this remnant resulted from a type-Ia supernova. Kepler is unusual among such remnants in that it shows shock-heated circumstellar material, indicating that its progenitor lost substantial mass before exploding. Thus, Kepler is a rare example of a 'prompt' type-Ia supernova, which exploded about 10^8 years after formation of its progenitor star[49]. This would require a progenitor star with a mass of ~8 solar masses that shed ~7 solar masses before becoming a white dwarf.

Chandra and XMM-Newton have also provided data which indicate that SNRs are important sites for the acceleration of cosmic rays[50,51]. The ability to distinguish the X-ray spectra of forward and reverse shocks has shown that X-ray emission from the forward shock is probably non-thermal synchrotron radiation from relativistic electrons accelerated in the forward shock. Non-thermal X-ray emission requires acceleration of electrons up to tens of teraelectronvolts and suggests acceleration of ions to even higher energies. (See ref. 51 for a discussion of other possible sources of cosmic rays in the Galaxy.) Observations of rapidly brightening and fading X-ray knots in SNRs RX J1713.7-3946 and Cas A imply short radiative lifetimes, strong magnetic fields and extremely efficient acceleration of electrons, and presumably ions, to high energies. Further evidence for the acceleration of ions in SNRs comes from the observed temperature behind the forward shock, which in some remnants is lower than expected from standard shock-wave theory. This implies that a significant portion of the post-shock energy is going into the acceleration of ions and electrons[50].

The shell-type SNRs are powered by the supernova itself. In contrast, filled-type (plerionic) SNRs—such as the Crab nebula (Fig. 7)— are powered by relativistic winds from young pulsars and ultimately their stored rotational energy. Chandra and XMM-Newton observations of these pulsar-wind nebulae have discovered[12] axisymmetric winds with shock fronts, jets and elongated nebulae[13].

Supernova remnants and pulsar wind nebulae will fade away after several thousand years, but if a neutron star or black hole is part of a binary star system, it may become an X-ray source once again as it accretes matter from its companion star. A long Chandra observation of the galaxy M33 provided a detailed study of the first eclipsing stellar-mass black-hole X-ray binary, M33-X7. This system contains one of the most massive (15.7 ± 1.4 solar masses) stellar black holes known and an unusually massive (70 solar masses) companion star. The system's properties—a massive black hole in a 3.45-day orbit, separated from its companion by only 42 times the radius of our Sun—are difficult to explain using conventional models for the evolution of massive stars, unless massive stars have significantly slower mass loss than usually thought[52]. Assuming that the black hole's spin is approximately aligned with the binary-system orbital angular momentum, analysis of spectral data combined from both observatories gave a precise measurement of the black hole's spin parameter[53]. In that an astrophysical black hole (being uncharged) can be described by just the mass and spin, this result yields a complete description of an asteroid-size object at a distance of about 3 million light years.
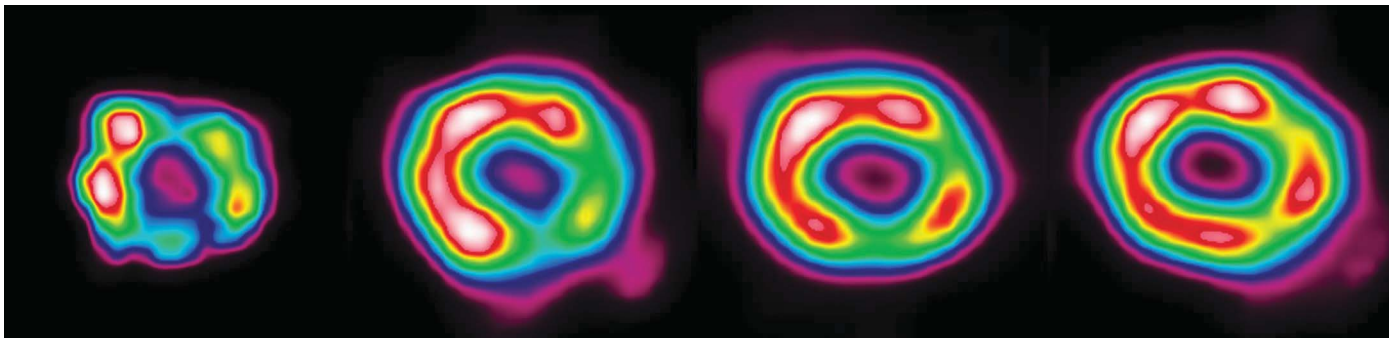

Figure 6 | Chandra (2″ × 2″) images detailing the expanding and brightening ring surrounding the stellar explosion SN 1987A. From left to right: January 2000, December 2002, August 2005 and October 2008. Reproduced with permission from K. Kowal Arcand and S. Park.
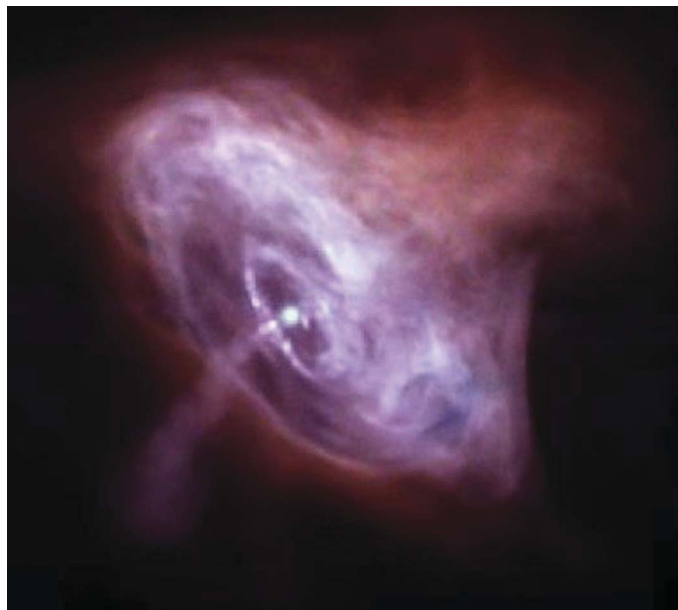
Figure 7 | Chandra (2′ × 2′) image of the Crab nebula and pulsar. The colours indicate intensity. Chandra provided the clearest view yet of the entire synchrotron nebula, discovering the inner ring and bright spot near the pulsar. Reproduced with permission from K. Kowal Arcand.

Supermassive black holes

Supermassive black hole census and evolution. On a grander scale than the collapse of a single star, a supermassive black hole forms during the course of evolution of a galaxy through as yet poorly understood processes. A considerable body of evidence suggests that most (if not all) massive galaxies harbour a supermassive black hole. The supermassive black hole's mass scales with the mass of the host-galaxy bulge, indicating that co-evolution of the galaxy and its super-massive black hole occurs[14]. The supermassive black hole in the centre of our Galaxy, coinciding with the radio source Sagittarius A*, has a mass of only $\sim 3 \times 10^6$ solar masses and is extremely faint outside the radio band. However, Chandra and XMM-Newton have detected weak X-ray emission from Sagittarius A* marked with frequent, rapid X-ray flaring[54,55], which presumably results from clumps of material falling towards the supermassive black hole.

A supermassive black hole can grow through the gradual accretion of ambient material or by episodic swallowing of large clumps of matter—including stars and other supermassive black holes. For the galaxy RX J1242.6-1119A, combined X-ray (ROSAT (Röntgensatellit), XMM-Newton and Chandra) observations and optical and ultraviolet data showed that tidal disruption scenarios, whereby a nearby star is captured by the supermassive black hole, can explain the observed X-ray flaring over months or even years[56]. The Chandra image of NGC 6240, an ultraluminous infrared galaxy that resulted from the collision of two smaller galaxies, revealed that the galaxy's central region contains two active supermassive black holes that are destined to merge in a few 100 million years. This discovery provides direct evidence for a key stage of galaxy mergers and the growth of super-massive black holes[57].

Gravitational energy released through accretion into a supermassive black hole is generally thought to power active galactic nuclei (AGNs), which exhibit a wide range of luminosities from radio and Seyfert galaxies to quasars. Because most AGNs produce strong X-ray radiation that can penetrate the clouds of dust and gas shrouding many supermassive black holes, Chandra and XMM-Newton have proven to be valuable tools to find and study such systems.

Through deep exposures, these observatories are tracing the evolution of the supermassive black hole population, thus revealing the accretion history of the Universe[9,58]. Wide, moderately deep surveys have enabled astronomers to find elusive, highly obscured AGNs, and

to study the population's evolution[59,60]. Figure 8 shows the XMM-Newton view of the Cosmic Evolution survey (COSMOS) field which is the deepest wide (>1 deg^2) extragalactic X-ray survey performed so far[61].

Among the most important findings is a strong cosmological evolution in the X-ray luminosity of AGNs[9,62–64]. Chandra and XMM-Newton data have yielded the first reliable space densities for X-ray-selected AGNs, covering a broad range in luminosity and redshift. The redshift measures the distance to an object and age of the Universe when the light was emitted: large redshifts imply large distances and young ages.

The evolutionary behaviour of AGNs depends upon X-ray luminosity: although the space density of high-luminosity AGNs peaks around a redshift $z \approx 2$, that of low-luminosity AGNs peaks at $z < 1$, confirming tendencies previously found with another X-ray satellite—ROSAT[62]. This drop occurs not because the accretion rate onto supermassive black holes evolves but rather because lower-mass supermassive black holes form or turn on more slowly than higher-mass supermassive black holes and because the latter switch off between $z = 1$ and the present. An explanation for this 'downsizing' is that high-luminosity AGNs are triggered in major mergers, whereas low-luminosity AGNs are due to more common non-merger events[65].

Combined X-ray spectral and timing observations also provide a probe of conditions near the event horizon of the supermassive black hole. XMM-Newton observations of the supermassive black hole in the galaxy MCG-6-30-15 have established that the broad iron line emission features observed can only be plausibly explained as lines shaped by relativistic effects[66]. In the galaxy NGC 3516, XMM-Newton observations of the modulation of a transient, redshifted iron $K\alpha$ emission line were used to track the orbit of a flaring hot spot just outside the event horizon for a supermassive black hole with a mass ~ 30 million solar masses[67]. The combined spectral and timing data for iron L-line emission from the Seyfert galaxy 1H0707-495 also provides strong evidence that we are witnessing emission from matter
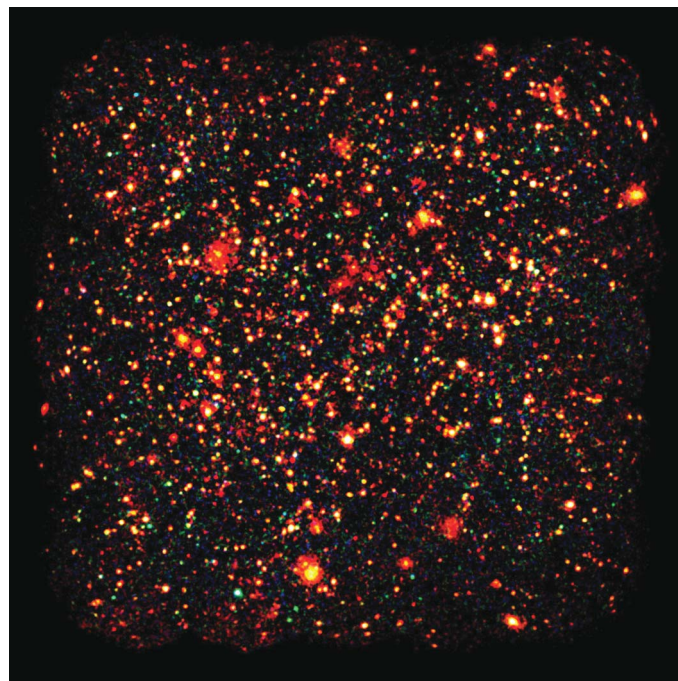

Figure 8 | X-ray image of thousands of sources in the XMM-Newton COSMOS field. The image covers three energy bands: 0.5–2 keV (red); 2–4.5 keV (green); and 4.5–10 keV (blue). Most of the $\sim$2,000 point-like sources are AGNs; more than 100 galaxy clusters appear extended. Reproduced with permission from G. Hasinger, N. Cappelluti and the XMM–COSMOS collaboration, and ESA.

within a fraction of a light minute from the event horizon of a rapidly spinning, massive black hole[68].

Supermassive black holes and clusters of galaxies. Accretion onto a black hole is typically characterized by episodic flaring activity and possible quasi-periodic oscillations[69]. Accretion can also be accompanied by, or in some cases replaced by, conical winds of gas photo-ionized by radiation from the AGN, extreme outflows, and powerful, relativistic jets (for example, Fig. 9), which can represent a substantial fraction of the energy budget of an AGN[14,70]. Supermassive-black-hole-driven jets[71,72] seem to be the primary mechanism by which the supermassive black hole interacts with the gas in its host galaxy.

Direct evidence of AGN feedback comes from Chandra and XMM-Newton observations of clusters of galaxies, which have diameters as large as 10 million light years. In the absence of a heat source, the plasma would exhibit a 'cooling flow' with increasingly higher density and lower temperature in the central regions. Cooling-flow models that ignored heating experienced wide popularity with few dissenters for almost two decades. The cooling-flow debate took a dramatic turn when XMM-Newton and Chandra spectra of the cores of a number of 'cooling flow' clusters found that emission lines from iron L-shell ions—particularly Fe XVII—were much fainter than predicted by the models. This sets a strong upper limit on the amount of plasma with temperatures less than 10 MK (ref. 73) and indicates the presence of a heat source.

Evidence for a heat source was provided by a Chandra image[74] of the Perseus cluster of galaxies. Two vast X-ray-faint cavities discovered a decade earlier with ROSAT[75] were mapped in exquisite detail by Chandra. Radio data[75] show that these X-ray holes are filled with magnetized bubbles of (relativistic) synchrotron-emitting electrons, which displace the hot thermal plasma as they expand and rise buoyantly from the central black hole. Chandra images show quasi-spherical, ripple-like structures that, if interpreted as sound waves, can provide sufficient energy to offset the cooling.

A number of other galaxy clusters show this type of activity[76], most notably the relatively nearby Virgo cluster of galaxies, where the X-ray images reveal intricate details of supermassive-black-hole-generated

outflow interacting with hot plasma[77,78]. The more distant cluster MS0735.6+7421 has huge X-ray cavities coincident with radio emission[76]. These cavities, which are 600,000 light years in diameter, indicate an outburst that has persisted for 100 million years and released energy of the order 10^{62} erg and indicate growth of the central black hole by as much as 100 million solar masses over that time.

The XMM-Newton and Chandra observations provide compelling evidence for a feedback scenario: gas cools in the cluster core and falls towards a central black hole. Accretion of the gas onto a spinning black hole then launches jets of relativistic particles that heat the cooling gas and stop the inflow, thereby shutting off the jets and allowing the cooling to resume. This picture also suggests an explanation for limits on the growth of massive galaxies at the centre of clusters[14].

Chandra and XMM-Newton have also provided new insights into the enrichment of the intracluster medium via supernova explosions. Many of the metals (elements from carbon to iron) in the intracluster medium are detected via their X-ray spectral signatures. Such data allow one to constrain models for supernova explosions[79] and to trace the iron content in clusters over cosmic time[80] with a decrease of nearly a factor of two from the present epoch to $z \approx 1$ when the Universe was about half its present age.

Dark matter and dark energy

During the past decade there has been great progress in identifying and measuring the mass-energy components of the Universe. On the basis of a broad suite of astronomical observations, a concordance model has emerged in which only about 4% of the Universe is normal (baryonic) matter, 23% is dark matter, and 73% is dark energy[81]. X-ray observations have had an important role in compiling evidence for this conclusion.

In the merger cluster 1E 0657-558 (see Fig. 10) Chandra observations showed that the distribution of X-ray-emitting plasma is different from that of the total mass. This provides direct evidence for the existence of dark matter, thus rejecting a modification of the gravitational-force law[15]. The huge dark-matter accumulations in clusters of galaxies are laboratories for studying dark-matter particles. For example, the absence of decay lines in the XMM-Newton X-ray spectra of the
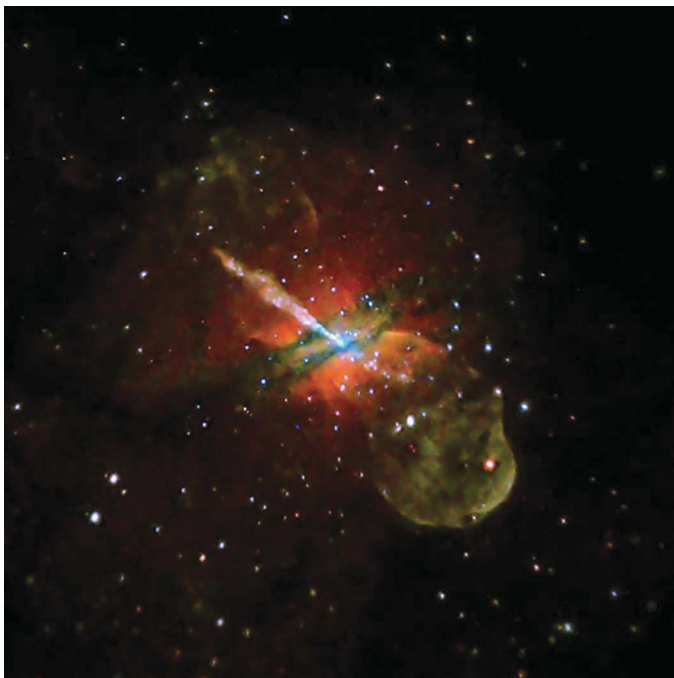

Figure 9 | Chandra image of Centaurus A. An ultra-deep X-ray view of the nearest galaxy with an active supermassive black hole. A prominent X-ray jet extends 13,000 light years to the upper left, with a shorter 'counterjet' in the opposite direction. Reproduced with permission of R. Kraft, NASA, the Chandra X-ray Center and the Center for Astrophysics.
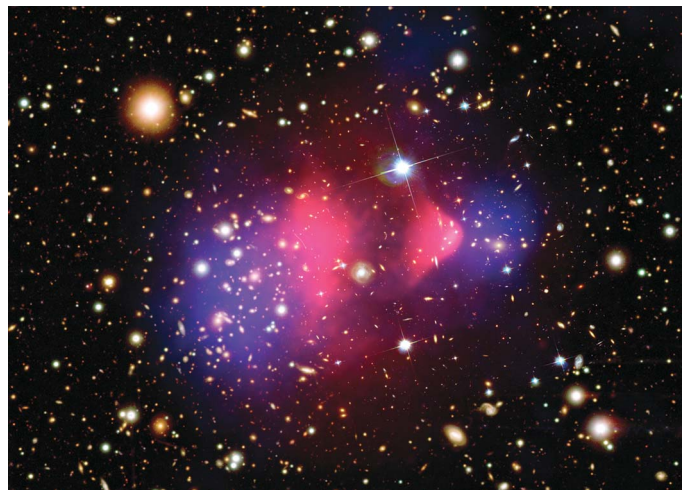

Figure 10 | Colliding clusters of galaxies and dark matter. Composite image of 1E0657-56 documents the collision of two large clusters of galaxies. Chandra resolved the X-ray emission (pink) produced by the vast clouds of hot gas which constitutes most of the baryonic matter. The optical image from Magellan and the Hubble Space Telescope shows the galaxies in orange and white. The blue areas indicate where most of the mass of the clusters lies, based upon gravitational weak lensing. Most of the matter (blue) is clearly separate from the normal matter (pink), directly showing that the mass in clusters is predominantly in some dark form that interacts only through gravity. Reproduced with permission from NASA, CXC and ref. 15.

Coma and Virgo clusters restricts the mass and mixing angle of hypothetical dark-matter sterile neutrinos[82].

Chandra and XMM-Newton data also showed that clusters of galaxies can be described by a universal mass profile[83,84]. This discovery constrains the dark-matter equation of state: self-interacting and adiabatic contracting dark matter can be eliminated[85,86].

On a larger scale, both the Hubble Space Telescope and XMM-Newton observed a 1.6-deg² region to map the three-dimensional structure of dark and baryonic matter. This large-scale structure comprises a loose network of filaments that grows over cosmic time. X-ray-emitting clusters of galaxies trace dark-matter concentrations at the intersections of these filaments[87].

Chandra observations of clusters of galaxies also provide a basis for precise determination of cosmological parameters in several new and unique ways. For example, the measured gas fraction f_{gas} (where f_{gas} = (mass in the X-ray-gas)/(total mass, including dark matter)) in clusters of galaxies depends on the assumed distance to a cluster, which in turn depends on the cosmological model. Moreover, for the largest, dynamically relaxed clusters, hydrodynamic simulations indicate that f_{gas} should be nearly constant with redshift. Combining Chandra observations of f_{gas} with this requirement yields constraints on the dark energy density and equation of state consistent and complementary with results obtained by other methods[16].

An independent technique for studying dark energy relies upon observing how the number of massive clusters per unit volume changes with time, compared with the predictions of cosmological models. With this 'growth of structure' method, Chandra data yield tight constraints on dark-matter and dark-energy content and on the dark-energy equation of state[17]. The data are consistent with the concordance cosmological model[81] wherein the dark energy arises from Einstein's cosmological constant.

Future prospects

We cannot predict with certainty the major areas of discovery in the future and the impact that they will have. However, on the premise that "what's past is prologue"[88], we are convinced that deeper and wider surveys will yield a wealth of new discoveries. We expect that future investigations will address critical issues ranging from the sub-nuclear composition of a neutron star to the large-scale structure of the Universe. Targets of interest extend from the geocorona to deep surveys providing a complete census and evolutionary history of AGNs out to redshifts $z \approx 7$. Extended surveys of star clusters in the Galaxy and deep field observations of distant galaxies will provide insights into the processes that trigger star formation, and how bursts of star formation in the Galaxy as well as in the early Universe enrich the ambient gas with heavy elements and feed energy back into the gas. Similarly, long observations of individual galaxy clusters will probe and refine our understanding of the mechanism by which the energy from supermassive black hole feedback is coupled to the intracluster gas.

The scientific return from XMM-Newton and Chandra observations will continue to be enhanced by collaborations with ground-based and other space telescopes. For example, combined X-ray, infrared and millimetre spectral diagnostics using Herschel and the Atacama Large Millimeter Array will be crucial for probing accretion and outflow processes (and the role of magnetic fields) in proto-stars, young stars and proto-planetary disks. The search for the warm, hot, intergalactic medium, via absorption lines in X-ray grating spectra of relatively X-ray-bright blazars, will select targets based on data from optical surveys. The X-ray observations may confirm existing tentative detections or set limits on the amount of intergalactic highly ionized oxygen, which traces the 'missing baryons' in the local Universe. Dark-energy parameters will be better constrained by X-ray observations of galaxy clusters at higher redshifts, including those discovered in recently initiated microwave surveys—for example, the South Pole Telescope and the Atacama Cosmology Telescope, as well as the Planck mission. A key goal of the X-ray

dark-energy research programme will be to test for a possible failure of general relativity through comparison of the effect of dark energy on cosmic expansion with its effect on the growth of cosmological structures.

1. Weisskopf, M. C. et al. An overview of the performance and scientific results from the Chandra X-Ray observatory. Publ. Astron. Soc. Pacif. 114, 1–24 (2002).
2. Jansen, F. et al. XMM-Newton observatory. I. The spacecraft and operations. Astron. Astrophys. 365, L1–L6 (2001).
3. Bhardwaj, A. et al. X-rays from solar system objects. Planet. Space Sci. 55, 1135–1189 (2007).
   Excellent overview of the progress made in this sub-discipline.
4. Guedel, M. & Naze, Y. X-ray spectroscopy of stars. Astron. Astrophys. Rev. 17, 309–408 (2009).
   Comprehensive review of the progress in stellar astronomy fostered by Chandra and XMM-Newton.
5. Wolk, S. J. et al. Stellar activity on the young suns of Orion: COUP observations of K5–7 pre-main-sequence stars. Astrophys. J. 160 (Suppl.), 423–449 (2005).
6. Vink, J. X-ray high resolution and imaging spectroscopy of supernova remnants. Proc. X-ray Universe 2005 (ESA SP-604) (ed. Wilson, A.) 319–327 (European Space Agency, 2006).
7. Weisskopf, M. C. & Hughes, J. P. Six Years of Chandra Observations of Supernova Remnants, Astrophysics Update 2 (ed. Mason, J. W.) 55–113 (Springer, 2006).
8. Revnivtsev, M. et al. Discrete sources as the origin of the Galactic X-ray ridge emission. Nature 458, 1142–1144 (2009).
9. Brandt, W. N. & Hasinger, G. Deep extragalactic X-ray surveys. Annu. Rev. Astron. Astrophys. 43, 827–859 (2005).
10. Turner, T. J. & Miller, L. X-ray absorption and reflection in active galactic nuclei. Astron. Astrophys. Rev. 17, 47–104 (2009).
11. Miller, J. M. et al. Stellar-mass black hole spin constraints from disk reflection and continuum modeling. Astrophys. J. 697, 900–912 (2009).
12. Weisskopf, M. C. et al. Discovery of spatial and spectral structure in the X-ray emission from the Crab nebula. Astrophys. J. 536, L81–L84 (2000).
    The clearest view yet of the entire X-ray synchrotron nebula, revealing the inner ring and bright spot near the pulsar.
13. Gaensler, B. M. & Slane, P. O. The evolution and structure of pulsar wind nebulae. Annu. Rev. Astron. Astrophys. 44, 17–47 (2006).
14. Cattaneo, A. et al. The role of black holes in galaxy formation and evolution. Nature 460, 213–219 (2009).
15. Clowe, D. et al. A direct empirical proof of the existence of dark matter. Astrophys. J. 648, L109–L113 (2006).
    Observations show that the distribution of X-ray-emitting plasma separated from that of the total mass, providing direct evidence for the existence of dark matter.
16. Allen, S. W. et al. Improved constraints on dark energy from Chandra X-ray observations of the largest relaxed galaxy clusters. Mon. Not. R. Astron. Soc. 383, 879–896 (2008).
17. Vikhlinin, A. et al. Chandra Cluster Cosmology Project III: Cosmological parameter constraints. Astrophys. J. 692, 1060–1074 (2009).
    Uses observed rate of growth of galaxy clusters to measure the accelerated expansion of the Universe.
18. Dennerl, K. X-rays from Venus observed with Chandra. Planet. Space Sci. 56, 1414–1423 (2008).
19. Bhardwaj, A. et al. First terrestrial soft X-ray auroral observation by the Chandra X-ray Observatory. J. Atmos. Sol. Terr. Phys. 69, 179–187 (2007).
20. Wargelin, B. J. et al. Chandra observations of the "dark" moon and geocoronal solar wind charge transfer. Astrophys. J. 607, 596–610 (2004).
21. Dennerl, K. X-Rays from Mars. Space Sci. Rev. 126, 403–433 (2006).
22. Branduardi-Raymont, G. et al. Spectral morphology of the X-ray emission from Jupiter's aurorae. J. Geophys. Res. 113, A02202 (2008).
23. Elsner, R. F. et al. X-ray probes of magnetospheric interactions with Jupiter's auroral zones, the Galilean satellites, and the Io plasma torus. Icarus 178, 417–428 (2005).
24. Branduardi-Raymont, G. Latest results on Jovian disk X-rays from XMM-Newton. Planet. Space Sci. 55, 1126–1134 (2007).
25. Ness, J. U., Schmitt, J. H. M. M. & Robrade, J. Detection of Saturnian X-ray emission with XMM-Newton. Astron. Astrophys. 414, L49–L52 (2004).
26. Ness, J. U. et al. X-ray emission from Saturn. Astron. Astrophys. 418, 337–345 (2004).
27. Bhardwaj, A. et al. Chandra observation of an X-ray flare at Saturn: evidence of direct solar control on Saturn's disk X-ray emissions. Astrophys. J. 624, L121–L124 (2005).
28. Bhardwaj, A. et al. The discovery of oxygen Kα X-ray emission from the rings of Saturn. Astrophys. J. 627, 73–76 (2005).
29. Bodewits, D. et al. Spectral analysis of the Chandra comet survey. Astron. Astrophys. 469, 1183–1195 (2007).
30. Schulz, R. et al. Detection of water ice grains after the DEEP IMPACT onto Comet 9P/Tempel 1. Astron. Astrophys. 448, L53–L56 (2006).
31. Snowden, S. L. et al. Observation of solar wind charge exchange emission from exospheric material in and outside Earth's magnetosheath 2008 September 25. Astrophys. J. 691, 372–381 (2009).

1003

32. Snowden, S. L., Collier, M. R. & Kuntz, K. D. XMM-Newton Observation of Solar Wind Charge Exchange Emission. *Astrophys. J.* **610**, 1182–1190 (2004).

33. Koutroumpa, D. *et al.* The solar wind charge-exchange contribution to the local soft X-ray background. Model to data comparison in the 0.1–1.0 keV band. *Space Sci. Rev.* **143**, 217–230 (2009).

34. Guedel, M. *et al.* The XMM-Newton extended survey of the Taurus molecular cloud (XEST). *Astron. Astrophys.* **468**, 353–377 (2007).

35. Feigelson, E. D. *et al.* Global X-ray properties of the Orion nebula region. *Astrophys. J.* **160** (Suppl.), 379–389 (2005).

36. Getman, K. V. *et al.* Chandra Orion Ultradeep Project: observations and source lists. *Astrophys. J.* **160** (Suppl.), 319–352 (2005).

37. Guedel, M. & Telleschi, A. The X-ray soft excess in classical T Tauri stars. *Astron. Astrophys.* **474**, L25–L28 (2007).

38. Tsujimoto, M. *et al.* Iron fluorescent line emission from young stellar objects in the Orion nebula. *Astrophys. J.* **160** (Suppl.), 503–510 (2005).

39. Favata, F. *et al.* Bright X-ray flares in Orion young stars from COUP: Evidence for star-disk magnetic fields? *Astrophys. J.* **160** (Suppl.), 469–502 (2005).

40. Alexander, R. From discs to planetesimals: Evolution of gas and dust discs. *N. Astron. Rev.* **52**, 60–77 (2008).

41. Guedel, M. *et al.* Million-degree plasma pervading the extended Orion nebula. *Science* **319**, 309–312 (2008).

42. Castor, J., McCray, R. & Weaver, R. Interstellar bubbles. *Astrophys. J.* **200**, L107–L110 (1975).

43. Zhekov, S. A. *et al.* Chandra LETG observations of supernova remnant 1987A. *Astrophys. J.* **645**, 293–302 (2006).

44. Zhekov, S. *et al.* High-resolution X-ray spectroscopy of SNR 1987A: Chandra LETG and HETG observations in 2007. *Astrophys. J.* **692**, 1190–1204 (2009).

45. Racusin, J. *et al.* X-ray evolution of SNR 1987A: The radial expansion. *Astrophys. J.* **703**, 1752–1759 (2009).

46. Laming, J. M. *et al.* The polar regions of Cassiopeia A: The aftermath of a gamma-ray burst? *Astrophys. J.* **644**, 260–273 (2006).

47. Lopez, L. A. *et al.* Tools for dissecting supernova remnants observed with Chandra: Methods and application to the galactic remnant W49B. *Astrophys. J.* **691**, 875–893 (2009).

48. Ho, W. & Heinke, C. A neutron star with a carbon atmosphere in the Cassiopeia A supernova remnant. *Nature* **462**, 71–73 (2009).

49. Reynolds, S. P. *et al.* A deep Chandra observation of Kepler's supernova remnant: A type Ia event with circumstellar interaction. *Astrophys. J.* **668**, L135–L138 (2007).

50. Reynolds, S. Supernova remnants at high energy. *Annu. Rev. Astron. Astrophys.* **46**, 89–126 (2008).

51. Butt, Y. Beyond the myth of the supernova-remnant origin of cosmic rays. *Nature* **460**, 701–704 (2009).

52. Abuberkerov, M. *et al.* The Mass of the black hole in the X-ray binary M33 X-7 and the evolutionary status of M33 X-7 and IC 10 X-1. *Astron. Rep.* **53**, 232–242 (2009).

53. Liu, J. *et al.* Precise measurement of the spin parameter of the stellar-mass black hole M33 X-7. *Astrophys. J.* **679**, L37–L40 (2008).

54. Baganoff, F. K. *et al.* Rapid X-ray flaring from the direction of the supermassive black hole at the Galactic Centre. *Nature* **413**, 45–48 (2001).

55. Porquet, D. *et al.* XMM-Newton observation of the brightest X-ray flare detected so far from Sgr A*. *Astron. Astrophys.* **407**, L17–L20 (2003).

56. Komossa, S. *et al.* A huge drop in the X-ray luminosity of the nonactive galaxy RX J1242.6–1119A, and the first postflare spectrum: testing the tidal disruption scenario. *Astrophys. J.* **603**, L17–L20 (2004).

57. Komossa, S. *et al.* Discovery of a binary active galactic nucleus in the ultraluminous infrared galaxy NGC 6240 using Chandra. *Astrophys. J.* **582**, L15–L19 (2003).

58. Brunner, H. *et al.* XMM-Newton observations of the Lockman Hole: X-ray source catalogue and number counts. *Astron. Astrophys.* **479**, 283–300 (2008). **Tracing the evolution of the supermassive black hole population, thus revealing the accretion history of the Universe.**

59. Fiore, F. *et al.* Chasing highly obscured QSOs in the COSMOS field. *Astrophys. J.* **693**, 447–462 (2009).

60. Brusa, M. *et al.* High-redshift quasars in the COSMOS survey: the space density of $z > 3$ X-ray selected QSOs. *Astrophys. J.* **693**, 8–22 (2009).

61. Hasinger, G. *et al.* The XMM-Newton wide-field survey in the COSMOS field: I. Survey description. *Astrophys. J.* **172** (Suppl.), 29–37 (2007).

62. Silverman, J. D. *et al.* The luminosity function of X-ray-selected active galactic nuclei: Evolution of supermassive black holes at high redshift. *Astrophys. J.* **679**, 118–139 (2008).

63. Yencho, B. *et al.* The OPTX Project. II. Hard X-ray luminosity functions of active galactic nuclei for $z \sim 5$. *Astrophys. J.* **698**, 380–396 (2009).

64. Hasinger, G., Miyaji, T. & Schmidt, M. Luminosity-dependent evolution of soft X-ray selected AGN. New Chandra and XMM-Newton surveys. *Astron. Astrophys.* **441**, 417–434 (2005).

65. Hopkins, P. & Hernquist, L. A characteristic division between the fueling of quasars and Seyferts: five simple tests. *Astrophys. J.* **694**, 599–609 (2009).

66. Miller, J. M. Relativistic X-ray lines from the inner accretion disks around black holes. *Annu. Rev. Astron. Astrophys.* **45**, 441–447 (2007).

67. Iwasawa, K., Miniutti, G. & Fabian, A. Flux and energy modulation of redshifted iron emission in NGC 3516: implications for the black hole mass. *Mon. Not. R. Astron. Soc.* **355**, 1073–1079 (2004).

68. Fabian, A. C. *et al.* Broad line emission from iron K- and L-shell transitions in the active galaxy 1H0707–495. *Nature* **459**, 540–542 (2009).

69. Gierliński, M. *et al.* A periodicity of ~1 hour in X-ray emission from the active galaxy RE J1034+396. *Nature* **455**, 369–371 (2008).

70. Kinkhabwala, A. *et al.* XMM-Newton reflection grating spectrometer observations of discrete soft X-ray emission features from NGC 1068. *Astrophys. J.* **575**, 732–746 (2002).

71. Worrall, D. M. The X-ray jets of active galaxies. *Astron. Astrophys. Rev.* **17**, 1–46 (2009).

72. Harris, D. & Krawczynsk, H. X-ray emission from extragalactic jets. *Annu. Rev. Astron. Astrophys.* **44**, 463–508 (2006).

73. Peterson, J. R. & Fabian, A. C. X-ray spectroscopy of cooling clusters. *Phys. Rep.* **427**, 1–39 (2006).

74. Fabian, A. C. *et al.* A deep Chandra observation of the Perseus cluster: shocks, ripples and conduction. *Mon. Not. R. Astron. Soc.* **344**, L43 (2003). **X-ray and radio structures in cluster core suggest that sound waves powered by jets from the central supermassive black hole provide the energy needed to offset radiative cooling losses.**

75. Boehringer, H. *et al.* A ROSAT HRI study of the interaction of the X-ray-emitting gas and radio lobes of NGC 1275. *Mon. Not. R. Astron. Soc.* **264**, L25–L28 (1993).

76. McNamara, B. & Nulsen, P. Heating hot atmospheres with active galacti nuclei. *Annu. Rev. Astron. Astrophys.* **45**, 117–175 (2007).

77. Forman, W. *et al.* Filaments, bubbles, and weak shocks in the gaseous atmosphere of M87. *Astrophys. J.* **665**, 1057–1066 (2007).

78. Simionescu, A. *et al.* The gaseous atmosphere of M 87 seen with XMM-Newton. *Astron. Astrophys.* **465**, 749–758 (2007).

79. de Plaa, J. *et al.* Constraining supernova models using the hot gas in clusters of galaxies. *Astron. Astrophys.* **465**, 345–355 (2007).

80. Anderson, M. E. *et al.* Redshift evolution in the iron abundance of the intracluster medium. *Astrophys. J.* **698**, 317–323 (2009).

81. Komatsu, E. *et al.* Five-year Wilkinson microwave anisotropy probe. Observations: Cosmological interpretation. *Astrophys. J.* **180** (Suppl.), 330–376 (2009).

82. Boyarsky, A. *et al.* Restrictions on parameters of sterile neutrino dark matter from observations of galaxy clusters. *Phys. Rev. D.* **74**, 3506 (2006).

83. Pointecouteau, E., Arnaud, M. & Pratt, G. W. The structural and scaling properties of nearby galaxy clusters. I. The universal mass profile. *Astron. Astrophys.* **435**, 1–7 (2005).

84. Vikhlinin, A. *et al.* Chandra sample of nearby relaxed galaxy clusters: Mass, gas fraction, and mass-temperature relation. *Astrophys. J.* **640**, 691–709 (2006).

85. Lewis, A. D., Buote, D. A. & Stocke, J. T. Chandra observations of A2029: the dark matter profile down to below 0.01 r_{vir} in an unusually relaxed cluster. *Astrophys. J.* **586**, 135–142 (2003).

86. Zappacosta, L. The absence of adiabatic contraction of the radial dark matter profile in the galaxy cluster A2589. *Astrophys. J.* **650**, 777–790 (2006).

87. Massey, R. *et al.* Dark matter maps reveal cosmic scaffolding. *Nature* **445**, 286–290 (2007).

88. Shakespeare, W. *The Tempest* Act 2, Scene 1 (Oxford Univ. Press, 1914).

89. Hwang, U. *et al.* A million second Chandra view of Cassiopeia A. *Astrophys. J.* **615**, L117–L120 (2004).

90. Ranalli, P. *et al.* A deep X-ray observation of M82 with XMM-Newton. *Mon. Not. R. Astron. Soc.* **386**, 1460–1480 (2008).

Author Contributions All of the authors worked together to produce the manuscript. M.C.W. led the effort on Solar System objects; M.S.-L. on individual stars and star-forming regions; M.C.W. on supernova remnants and massive stellar black holes; N.S. on supermassive black hole census and evolution; H.T. and W.T. on supermassive black holes and clusters of galaxies; and N.S. on dark matter and dark energy. W.T. served most ably as editor and supported all the sections. M.C.W. was the executive editor.

Complex landscapes of somatic rearrangement in human breast cancer genomes

Philip J. Stephens[1], David J. McBride[1], Meng-Lay Lin[1], Ignacio Varela[1], Erin D. Pleasance[1], Jared T. Simpson[1], Lucy A. Stebbings[1], Catherine Leroy[1], Sarah Edkins[1], Laura J. Mudie[1], Chris D. Greenman[1], Mingming Jia[1], Calli Latimer[1], Jon W. Teague[1], King Wai Lau[1], John Burton[1], Michael A. Quail[1], Harold Swerdlow[1], Carol Churcher[1], Rachael Natrajan[2], Anieta M. Sieuwerts[3], John W. M. Martens[3], Daniel P. Silver[4], Anita Langerød[5], Hege E. G. Russnes[5], John A. Foekens[3], Jorge S. Reis-Filho[2], Laura van 't Veer[6], Andrea L. Richardson[4,7], Anne-Lise Børresen-Dale[5,8], Peter J. Campbell[1], P. Andrew Futreal[1] & Michael R. Stratton[1,9]

Multiple somatic rearrangements are often found in cancer genomes; however, the underlying processes of rearrangement and their contribution to cancer development are poorly characterized. Here we use a paired-end sequencing strategy to identify somatic rearrangements in breast cancer genomes. There are more rearrangements in some breast cancers than previously appreciated. Rearrangements are more frequent over gene footprints and most are intrachromosomal. Multiple rearrangement architectures are present, but tandem duplications are particularly common in some cancers, perhaps reflecting a specific defect in DNA maintenance. Short overlapping sequences at most rearrangement junctions indicate that these have been mediated by non-homologous end-joining DNA repair, although varying sequence patterns indicate that multiple processes of this type are operative. Several expressed in-frame fusion genes were identified but none was recurrent. The study provides a new perspective on cancer genomes, highlighting the diversity of somatic rearrangements and their potential contribution to cancer development.

Cytogenetic studies over several decades have shown that somatic rearrangements, in particular chromosomal translocations, occur in many human cancer genomes[1–3]. The prevalence of rearrangements is, however, variable with some cancer genomes exhibiting few and others, including the genomes of many common adult epithelial cancers, showing many.

Somatic rearrangement is a common mechanism for the conversion of normal genes into cancer genes[1–5]. Indeed, of the ~400 genes that are currently known to be somatically mutated and implicated in cancer development, most are altered by genomic rearrangement (http://www.sanger.ac.uk/genetics/CGP/Census/). These rearrangements usually result in the formation of a fusion gene, derived from two disrupted normal genes, from which a fusion transcript and protein is generated. In some instances, however, rearrangements place an intact gene under the control of new regulatory elements or cause internal reorganization of a gene. These events usually result in activation of the protein to contribute to oncogenesis, as in the paradigm of the *BCR–ABL* fusion gene in chronic myeloid leukaemia[6].

Most of the currently known fusion genes are operative in leukaemias, lymphomas and sarcomas[1,3], although similar cancer-causing rearrangements in *RET*, *NTRK1*, *NTRK3*, *BRAF* and *TFE3* were reported in rare epithelial cancers many years ago (http://www.sanger.ac.uk/genetics/CGP/Census/). Activated fusion genes in common adult epithelial cancers, such as those of the *ETS* transcription factor family in prostate cancer[7] and the *ALK* gene in lung adenocarcinoma[8],

were discovered only recently and not through the conventional strategy of positional cloning of cytogenetically ascertained chromosomal breakpoints. Their late emergence primarily reflects the complexity of cytogenetically visible rearrangement patterns in the genomes of many adult epithelial cancers and the consequent difficulty in prior selection of driver rearrangements for further study among the many likely passenger changes[9]. Rearrangements also constitute a subset of mutational events that result in inactivation of recessive cancer genes (tumour suppressor genes) and underlie genomic amplifications that result in increased copy number of cancer genes[10–12].

In recent years, the diversity of prevalence and pattern of point mutations and copy number changes in cancer genomes have been elucidated by systematic polymerase chain reaction (PCR)-based resequencing studies and by application of high-resolution copy number arrays, respectively[13–16]. Understanding of the basic patterns of rearrangement in most cancers, however, remains rudimentary. We recently demonstrated that second-generation sequencing of both ends of large numbers of DNA fragments generated from cancer genomes allows comprehensive characterization of rearrangements[17]. Here we apply this approach to a series of breast cancers to explore patterns of rearrangement and their potential contribution to cancer development.

Landscapes of rearrangement

Twenty-four breast cancers were investigated by sequencing both ends of ~65,000,000 randomly generated ~500-bp DNA fragments

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]Molecular Pathology Laboratory, The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK. [3]Department of Medical Oncology, Josephine Nefkens Institute, Erasmus University Medical Center, PO Box 2040, 3000 CA, Rotterdam, The Netherlands. [4]Department of Cancer Biology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Department of Genetics, Norwegian Radium Hospital, Oslo University Hospital, Montebello, N-0310 Oslo, Norway. [6]The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands. [7]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [8]Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Montebello, 0310 Oslo, Norway. [9]Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.

from each cancer on Illumina GAII Genome Analysers (Supplementary Fig. 1). The series included primary tumours and immortal cancer cell lines, examples of the commonest phenotypically defined subtypes and cases with high-risk germline predisposition mutations in *BRCA1*, *BRCA2* and *TP53* (Table 1). Rearrangements were initially identified as discordant paired-end reads that did not map back to the reference human genome in the correct orientation with respect to each other and/or within ~500 bp of each other. These were subsequently confirmed and evaluated for somatic or germline origin.

A total of 2,166 confirmed somatic rearrangements were identified among the 24 cancers (Supplementary Tables 1 and 2). The presence of multiple read pairs spanning each rearrangement (Supplementary Table 1), easily detectable copy number changes associated with many and targeted fluorescence *in situ* hybridization (FISH) studies on a subset indicate that most rearrangements are in cells of the dominant cancer clone and not in minority cell populations. By investigating whether rearrangements were found for 200 changes in genomic copy number, we estimate that ~50% of rearrangements have been detected.

Some cancers carried many rearrangements. For example, the breast cancer cell line HCC38 has at least 238 (Fig. 1), many more than could have been predicted from cytogenetic studies (http://www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/HCC38.html). However, there is substantial variation in prevalence, with some primary breast cancers carrying only a single rearrangement (Fig. 1, Supplementary Tables 1 and 2, and Supplementary Fig. 2). Overall, breast cancer cell lines showed more rearrangements (median 101, range 58–245) than primary cancers (median 38, range 1–231). This difference may be due to the acquisition of additional rearrangements during *in vitro* culture. However, it may also reflect less sensitive detection of rearrangements in primary cancers due to contaminating normal tissue or the relative propensity of some subclasses of breast cancer, for example metastases generating pleural effusions, to become established in culture. In some cancers rearrangements are evenly distributed through the genome. By contrast, in others they cluster in and connect genomic regions showing

amplification. The rearrangement architecture in such amplicons is often highly complex[10,11].

Architectures of rearrangement

The orientations and relative chromosomal locations of the two genomic segments forming each rearrangement can be used as the basis of a rearrangement classification system. This may be further elaborated using information from copy number and other analyses that allow reconstruction of the genomic architecture associated with each rearrangement. For the purposes of this report, we have derived a simplified version of this system, classifying each rearrangement according to (1) whether it is in an amplicon or not; (2) if not in an amplicon whether it is interchromosomal or intrachromosomal; and (3) if intrachromosomal whether it results in a deletion, tandem duplication, or rearrangement of inverted orientation (Fig. 1b).

There were 1,311 intrachromosomal and 239 interchromosomal rearrangements (excluding rearrangements within amplicons of which 397 were intrachromosomal compared with 219 interchromosomal) (Table 2 and Supplementary Tables 1 and 2). Thus, intrachromosomal rearrangements substantially outnumber interchromosomal rearrangements in breast cancer genomes. The breakpoints of 1,574 out of 1,708 intrachromosomal rearrangements were within 2 Mb of each other. These patterns are presumably attributable to the greater likelihood of physical interaction between two positions on the same chromosome compared to positions on different chromosomes—perhaps due to individual chromosomes occupying domains in the nucleus—and between two locations that are in close proximity on the same chromosome. The predominance of intrachromosomal rearrangements has not previously been appreciated because of the limited resolution of cytogenetic studies and because FISH-based approaches, such as spectral karyotyping, generally use a single fluorescence colour per chromosome.

The most commonly observed architecture of rearrangement was tandem duplication (there were 739 tandem duplications, 357 deletions and 215 rearrangements with inverted orientation; Table 2 and Supplementary Tables 1 and 2). The evidence that these are truly tandem insertions derives from the structure of the genomic rearrangement itself supported by complementary DNA and FISH studies (Tables 3 and 4 and Supplementary Figs 3 and 4). The duplicated segments ranged in size from 3 kb to greater than 1 Mb (Supplementary Table 1). Despite being the commonest class of rearrangement in breast cancer, tandem duplications have previously been overlooked because they are intrachromosomal and involve small chromosomal segments beyond the resolution of cytogenetics or previous generations of copy number arrays.

There were major differences between individual breast cancers in the number of tandem duplications (Fig. 1b and Supplementary Fig. 5). Some exhibited more than one-hundred whereas others showed few or none. The mechanistic basis for these differences is unknown, but may be due to a defective DNA maintenance process which confers a 'mutator phenotype'. This would be similar, in principle, to the defects in DNA mismatch repair that are responsible for microsatellite instability in some cancers. This putative mutator phenotype is unlikely to be directly related to deficiencies in BRCA1- or BRCA2-mediated DNA repair, as some cancers arising in individuals with germline mutations in these genes have few tandem duplications. Large numbers of tandem duplications were generally observed, however, in cancers that do not express oestrogen and progesterone receptors.

Sequences at rearrangement junctions

DNA sequence across the rearrangement junction was obtained from 1,821 (3,642 breakpoints) of the 2,166 confirmed rearrangements (Supplementary Table 3). The sequences 100 bp either side of each breakpoint were examined for the presence of motifs and sequence content. No striking signatures were observed, although there was a slight deficit of C•G base pairs compared to the genome as a whole ($P < 0.001$) and modest enrichment of some motifs (Supplementary

Table 1 | The 24 breast cancers analysed for somatic rearrangements

Sample name	Sample type	Age at diagnosis (years)	Grade	ER status	PR status	ERBB2 status	Germline mutations
HCC1143	Cell line	52	3	−	−	−	−
HCC1187	Cell line	41	3	−	−	−	−
HCC1395	Cell line	43	3	−	−	−	BRCA1
HCC1599	Cell line	44	3	−	−	−	−
HCC1937	Cell line	23	3	−	−	−	BRCA1
HCC1954	Cell line	61	3	−	−	+	−
HCC2157	Cell line	48	2	−	+	+	−
HCC2218	Cell line	38	3	−	−	+	TP53
HCC38	Cell line	50	3	−	−	−	−
PD3664a*	Primary	87	3	−	−	−	−
PD3665a*	Primary	47	3	−	−	−	−
PD3666a†	Primary	86	2	+	+	−	−
PD3667a†	Primary	43	2	+	+	−	−
PD3668a‡	Primary	72	2	+	+	−	−
PD3669a‡	Primary	66	2	+	+	−	−
PD3670a§	Primary	66	3	−	−	+	−
PD3671a§	Primary	48	3	−	−	+	−
PD3672a‖	Primary	78	2	−	−	−	−
PD3687a	Primary	39	3	−	−	−	BRCA1
PD3688a	Primary	47	3	−	−	−	BRCA1
PD3689a	Primary	50	2	+	+	−	BRCA2
PD3690a	Primary	40	2	+	+	−	BRCA2
PD3693a	Primary	48	3	−	−	+	−
PD3695a	Primary	66	3	−	−	−	−

All the breast cancer samples screened were invasive ductal carcinomas with the exception of PD3672a, which was an invasive lobular carcinoma. ER status refers to expression of oestrogen receptor; PR status refers to expression of progesterone receptor.
* Sample subset classified as basal-type according to expression profile.
† Sample subset classified as luminal A according to expression profile.
‡ Sample subset classified as luminal B according to expression profile.
§ Sample subset classified as ERBB2 according to expression profile.
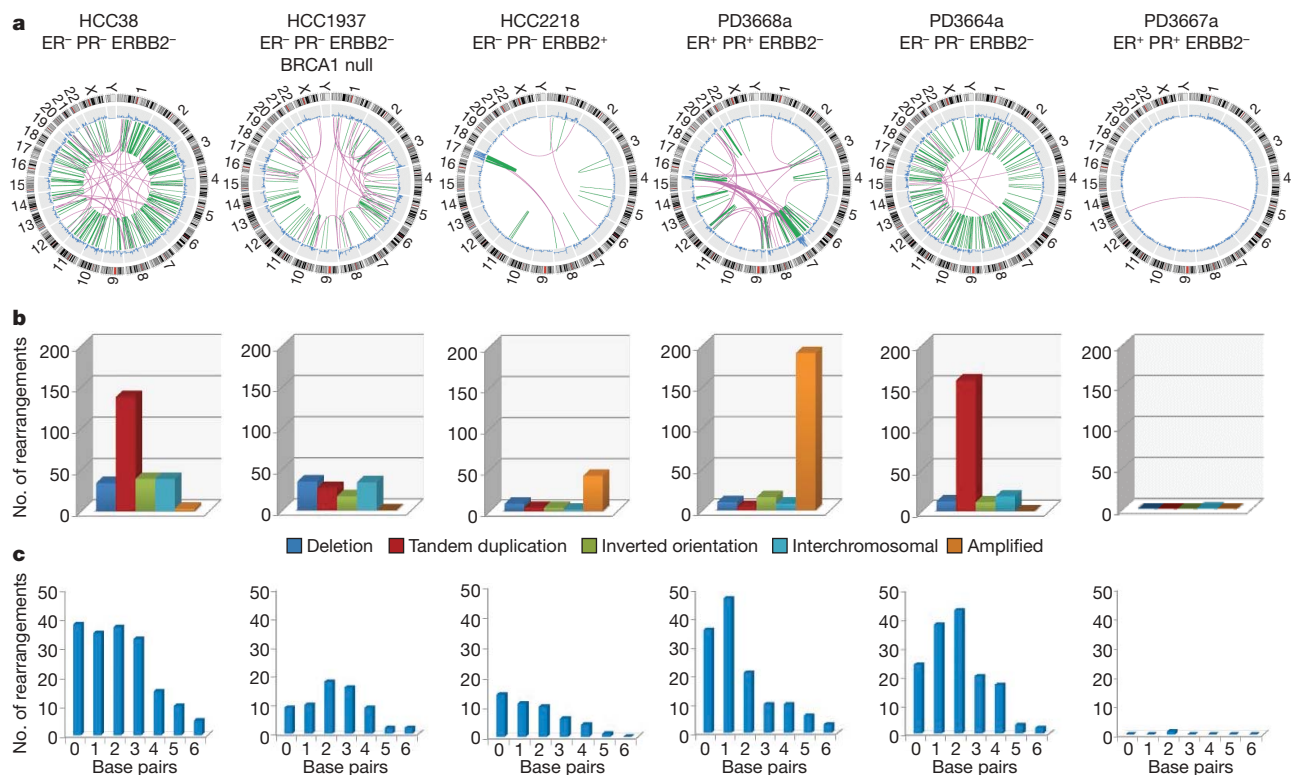‖ Sample subset classified as normal-like according to expression profile.

Figure 1 | Somatic rearrangements observed in six of the twenty-four breast cancer samples screened. a, Genome-wide Circos plots of somatic rearrangements. An ideogram of a normal karyotype is shown in the outer ring. A copy number plot is represented by the blue line shown inner to the chromosome ideogram. Within the inner ring each green line denotes an intrachromosomal rearrangement and each purple line an interchromosomal rearrangement. ER, oestrogen receptor; PR, progesterone receptor. **b,** The prevalence of rearrangement architectures in individual cancers: deletion (dark blue), tandem duplication (red), inverted orientation (green), interchromosomal rearrangements (light blue), rearrangements within amplified regions (orange). **c,** Extent of overlapping microhomology at rearrangement breakpoints. The number of base pairs of microhomology is plotted on the horizontal axis.

Tables 4 and 5). However, no single motif was commonly found in any class of rearrangement.

The sequences either side of each rearrangement junction were then compared to each other. In most instances the two contributing DNA segments showed a short stretch of identical sequence, known as an overlapping microhomology, immediately adjacent to the rearrangement junction which was present only once in the re-arranged DNA (Fig. 1c, Table 2, Supplementary Table 3 and Supplementary Fig. 6). Approximately 15% of rearrangements showed non-templated sequence at the rearrangement junction (Table 2, Supplementary Table 3 and Supplementary Fig. 7). In many, this is only a few base pairs long, although the longest segment of this type was 154 bp. A further 3.8% of rearrangements included one or more small fragments of DNA (<500bp) from elsewhere in the genome

interposed between the rearrangement breakpoints identified by the paired end sequencing. We have previously termed these small DNA fragments 'genomic shards'[10,17].

Overlapping microhomologies and non-templated sequences at rearrangement junctions are often considered to be signatures of a non-homologous end-joining (NHEJ) DNA double-strand break repair process[18–21]. The segments of overlapping microhomology

Table 2 | Summary of rearrangement patterns found in 24 breast cancers

Rearrangement class	Number in cell lines	Number in primaries	Total (%)
Deletion	214 (23.8 (9–35))	143 (9.5 (0–41))	357 (16.5)
Tandem duplication	370 (41.1 (4–138))	369 (24.6 (0–158))	739 (34)
Inverted orientation	113 (12.6 (4–24))	102 (6.8 (0–18))	215 (10)
Interchromosomal	147 (16.3 (2–39))	92 (6.1 (0–27))	239 (11)
Amplified	308 (34.2 (0–208))	308 (20.5 (0–191))	616 (28.5)
Total	1,152 (128 (58–245))	1,014 (67.6 (1–231))	2,166 (100)

Microhomology at rearrangement junctions (bp)		Non-templated sequence at rearrangement junctions (bp)	
Rearrangement class	Mean (range)	Rearrangement class	Mean (range)
Deletion	2.03 (0–14)	Deletion	3.27 (0–42)
Tandem duplication	2.10 (0–9)	Tandem duplication	3.46 (0–48)
Inverted orientation	2.50 (0–21)	Inverted orientation	5.04 (0–45)
Interchromosomal	2.00 (0–9)	Interchromosomal	3.63 (0–60)
Amplified	1.71 (0–9)	Amplified	3.83 (0–154)
Total	2.00 (0–21)	Total	3.71 (0–154)

In the top section of Table 2, the mean per case and range are indicated in parentheses.

Table 3 | Expressed in-frame fusion genes found in the 24 breast cancers

Sample name	5' gene	3' gene
HCC1187†	PLXND1	TMCC1
HCC1187	RGS22	SYCP1
HCC1395	EFTUD2	KIF18B
HCC1395	ERO1L	FERMT2
HCC1395*	KCNQ5	RIMS1
HCC1395	PLA2R1	RBMS1
HCC1599	CYTH1	PRPSAP1
HCC1937	NFIA	EHF
HCC1954	STRADB	NOP58
HCC2157	INTS4	GAB2
HCC2157	RASA2	ACPL2
HCC2157	SMYD3	ZNF695
HCC38	ACBD6	RRP15
HCC38	LDHC	SERGEF
HCC38	MBOAT2	PRKCE
HCC38	SLC26A6	PRKAR2A
HCC38	HMGXB3	PPARGC1B
PD3664a	RAF1	DAZL
PD3670a	AC141586.2	CCNF
PD3670a	SEPT8	AFF4
PD3688a	ETV6	ITPR2
PD3693a*	HN1	USH1G

Gene accession numbers and exons fused are outlined in Supplementary Table 6.
* Gene fusion is amplified.
† Predicted to be an out-of-frame gene fusion. However, RT–PCR across the exon–exon fusion boundary demonstrated both an out-of-frame and an in-frame gene fusion due to alternative splicing.

1007

Table 4 | Expressed in-frame rearranged genes found in the 24 breast cancers

Sample name	Genes
HCC1187	*F8, FBXL20, GMDS, MED13L, RB1*
HCC1395	*CADPS2, DYNC2H1, KIAA0802, MBTPS1, TLE1*
HCC1937	*LRBA, RUNX1, SEMA3D, SSBP3*
HCC38	*EPHA3, EPS15, FRY, KCNMB2, REPS1, SLC4A4, GIGYF2*
PD3664a	*C12orf35, FAM164A, GABRP, HOMER2, INADL, KCNMA1, NFE2L3, ODZ1, PDE4B, SVIL, VPS8*
PD3665a	*DAPK1*
PD3668a	*PLCB1, SYNJ1*
PD3671a	*KIAA0146*
PD3687a	*GP1, WRN*
PD3693a	*MACROD2*

Gene accession numbers and exons fused are outlined in Supplementary Table 7.

are believed to mediate alignment of the two DNA fragments that are joined. It has, however, recently been proposed that complex germ-line rearrangements with genomic shards and overlapping microhomology might be due to replicative mechanisms[21]. The small proportion of complex rearrangements with genomic shards may indicate that this mechanism is relatively infrequently operative in breast cancer.

It has previously been suggested that there exist multiple NHEJ repair processes that may be characterized by different lengths of overlapping microhomology at rearrangement junctions[21,22]. To investigate this possibility we examined the distribution of microhomology lengths in each breast cancer (Fig. 1c and Supplementary Fig. 6). In some breast cancers, rearrangements with zero base pairs of microhomology were most frequent, whereas in others rearrangements with two or more base pairs were the commonest class. Rearrangements with zero base pairs of microhomology were most common in amplicons, in contrast to all other classes of rearrangement in which the modal class of microhomology was 2 bp (Fig. 2). These differences are unlikely to be due to chance ($P < 0.001$) and indicate that there are at least two classes of NHEJ repair which are operative to different extents in different breast cancers[21].

Because the analysis of paired-end sequences requires alignment to the reference human genome and because sequences within repetitive elements are more likely to misalign it is conceivable that we have missed classes of rearrangement mediated by repeats. To investigate this possibility further we constructed libraries from the breast cancer cell line HCC1187 in which the sequenced ends were 3 kb rather than 500 bp apart. The 3-kb paired ends will flank the majority of common repeats and thus allow detection of rearrangements mediated by them. Although additional rearrangements were detected, a distinct class of repeat-mediated rearrangement was not found (data not shown).

Rearrangements of protein coding genes

Fifty per cent of rearrangements fell within the footprint of a protein coding gene compared to 40% expected by chance ($P < 10^{-7}$). The reasons for this striking enrichment of rearrangements in genic regions are not clear. Because rearrangements that confer selective advantage on a cancer clone are a priori more likely to be located in

genes it is conceivable that some of this effect is due to selection and that a subset of rearrangements is implicated in cancer development. However, it may be more plausible that there are structural properties of genic regions that increase the likelihood of a DNA double-strand break occurring, perhaps dependent on active transcription or chromatin configuration.

Twenty-nine rearrangements were predicted to generate in-frame gene fusions. Using exon–exon PCR with reverse transcription (RT–PCR), rearranged transcripts from 19 out of 22 in-frame fusion genes in non-amplified regions and from 2 out of 6 (1 not determined) in amplified regions were found (Table 3). Thus, most in-frame rearranged genes from non-amplified regions have the requisite 5′ and 3′ DNA sequences for transcript formation and stability. Conversely, most from amplified regions do not and these rearrangements probably represent fragments of one or both genes reflecting the high density of rearrangements often present in these regions[10]. Sixty-six in-frame internally rearranged genes were also identified. Rearranged transcripts were present in 39 out of 58 assessed (Table 4). In some cancers multiple in-frame rearranged and expressed genes are present (Tables 3 and 4 and Supplementary Tables 6 and 7).

Several in-frame fusion genes are potentially of biological interest as candidates for new cancer genes. Notably, two were members of the *ETS* family of transcription factors. *ETV6* is rearranged to form cancer genes with multiple different fusion partners in leukaemias[23], congenital fibrosarcoma[24] and myelodysplastic syndrome. It also forms a rearranged cancer gene with *NTRK3* in the rare subclass of secretory breast cancer[25]. Here, *ETV6* was fused to *ITPR2* (Fig. 3) through an inversion involving intron 2, a site previously reported in other cancers[23], and was rearranged in a further breast cancer without clearly forming an in-frame fusion gene. *ITPR2* encodes inositol-1,4,5-triphosphate receptor type 2 which is involved in signal transduction and regulation of cellular calcium fluxes. The second rearrangement fused *EHF*, which has not been previously implicated in cancer development, to *NFIA*, a transcription factor involved in adenovirus replication (Supplementary Fig. 3).

Fusion genes implicated in cancer development are likely to be recurrent. However, none of the novel fusion genes we identified was present in more than one out of the 24 cancers screened. Three expressed, in-frame fusion genes were examined by FISH (*ETV6–ITPR2*, *NFIA–EHF* and *SLC26A6–PRKAR2A*) and twenty by RT–PCR across the rearranged exon–exon junction in 288 additional breast cancer cases. No further examples were found, indicating that they are either passenger events or that they contribute infrequently to breast cancer development.

Rearrangements were found in several known cancer genes including *BRAF, PAX3, PAX5, NSD1, PBX1, MSI2* and *ETV6* (see above). Each of these is a partner in a fusion gene in other classes of human cancer and was rearranged in two of the 24 samples analysed, although in many cases an in-frame fusion gene was not obviously generated (Supplementary Table 8). Rearrangements found in *RB, APC, FBXW7* and other recessive cancer genes may have resulted in gene inactivation to contribute to cancer development.

Several other genes were rearranged in multiple cancers (Supplementary Table 9). Some are in amplified regions surrounding
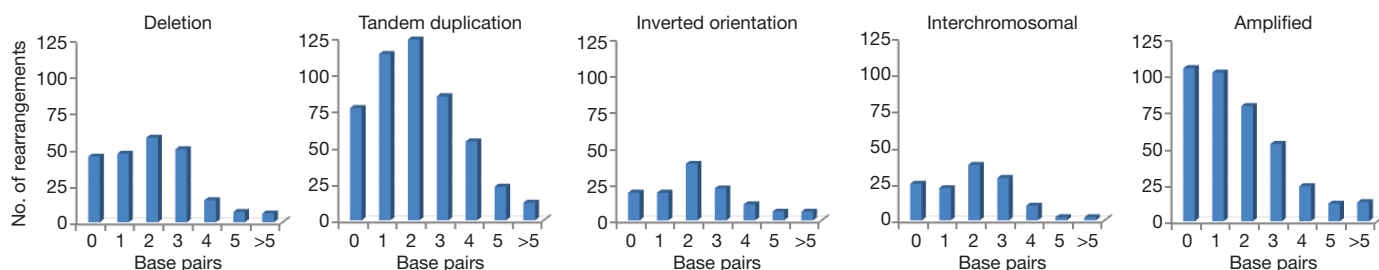

Figure 2 | Extent of overlapping microhomology at different architectural classes of rearrangement junctions. The number of base pairs of microhomology is plotted on the horizontal axis.
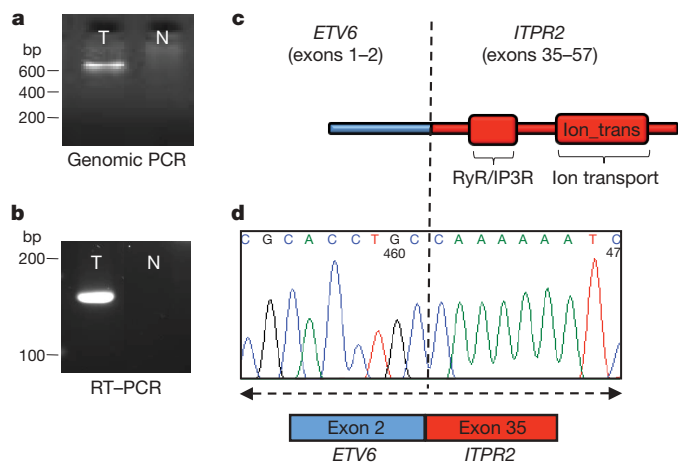
Figure 3 | ETV6–ITPR2, an expressed, in-frame fusion gene generated by a 15-Mb inversion in the primary breast cancer PD3668a. a, Across-rearrangement PCR to confirm the presence of the somatic rearrangement. **b,** RT–PCR of RNA between *ETV6* exon 2 and *ITPR* exon 35 to confirm the presence of a chimaeric expressed transcript. N, normal; T, tumour. **c,** Schematic diagram of the protein domains fused in the predicted *ETV6–ITPR2* fusion protein. RyR/IP3R, ryanodine receptor and inositol triphosphate receptor homology associated domain. **d,** Sequence from RT–PCR product shown in **b** confirming *ETV6* exon 2 fused to *ITPR2* exon 35.

ERBB2 (for example *ACCN1*, which is rearranged in four out of the 24 breast cancers) or other known targets of genomic amplification in breast cancer. It is likely that these are recurrently rearranged because of the high density of rearrangements associated with these regions of recurrent genomic amplification. Others, however, are not in regions of genomic amplification. For example, *SHANK2* was rearranged in five of the 24 breast cancers, whereas *IGF1R*, *GRHL2*, *EFNA5* and *MACROD2* were each rearranged in four. These recurrently rearranged genes generally have large genomic footprints and may simply represent bigger targets for randomly positioned rearrangements (Supplementary Table 9). For some, however, an elevated local rate of DNA double strand breakage ('fragility') may also contribute to the clustering of rearrangements.

Discussion

This study has generated the most comprehensive insight thus far into patterns of somatic rearrangement in cancer genomes. Most rearrangements in breast cancer are intrachromosomal. Tandem duplications seem to be the most common subclass and are known to form activated cancer genes in other cancer types[26,27]. The high prevalence of tandem duplications in a subset of cancers suggests the presence of a defect in DNA maintenance which generates this particular class of rearrangement. The underlying abnormality responsible for this phenotype is unknown. It may reside in the licensing mechanisms responsible for defining, priming and monitoring origins of DNA replication[28].

Breast cancers are highly heterogeneous and are subclassified on the basis of oestrogen receptor, progesterone receptor and *ERBB2* expression and by messenger RNA expression profiles[29,30]. Subclasses defined in these ways show correlations with patterns of genomic alteration[31,32]. Breast cancers with many tandem duplications are usually oestrogen- and progesterone-receptor negative and classified by expression profile as basal-like. In contrast, cancers with few rearrangements or with rearrangements within amplicons (other than those involving *ERBB2*) are usually oestrogen-receptor positive and classified as luminal A and luminal B types, respectively.

Many novel in-frame fusion genes or internally rearranged genes were identified, most of which were expressed. None, however, were found to be recurrent. Approximately 2% of rearrangements would be expected to generate an in-frame fusion gene by chance, compared to

1.6% observed. It is therefore likely that most are passenger events. Nevertheless, as previously suggested for somatic point mutations[13,14] it may be that multiple, infrequently rearranged cancer genes are operative in breast cancer as they are in leukaemia[2]. Furthermore, detailed analysis of rearrangement breakpoints will be necessary to investigate the possibility of fusions between promoters/regulatory elements and intact genes that result in deregulation of expression. Much larger series will be required to investigate comprehensively the possibility of recurrent cancer-causing rearrangements in breast cancer.

Exhaustive sequencing of substantial numbers of cancer genomes to yield complete catalogues of all classes of somatic mutations will gather pace over the next few years. The current study offers insight into the complexity of rearrangement patterns that will be encountered in solid tumour genomes, demonstrates the potential for generation of active rearranged genes that may be implicated in cancer development, and illustrates the types of information that will emerge on mutational processes that will have been operative during the development of individual cancers.

METHODS SUMMARY

Paired-end sequence reads of 37 bp were generated on the Illumina Genome Analyser from ~500-bp insert genomic libraries and aligned to the human reference genome (NCBI Build 36) using MAQ[33]. Correctly mapping reads were used to generate genome-wide copy number information and discordantly mapping reads defined potential rearrangements[17]. Putative rearrangements were PCR amplified in tumour and matching normal DNA and proven somatic rearrangements were capillary sequenced to resolve breakpoints to base-pair level. Structural variants predicted to generate novel gene fusions or to internally rearrange genes were further characterized by RT–PCR and sequencing from tumour-derived total RNA.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Kaye, F. J. Mutation-associated fusion cancer genes in solid tumors. *Mol. Cancer Ther.* **8,** 1399–1408 (2009).
2. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genet.* **36,** 331–334 (2004).
3. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nature Rev. Cancer* **7,** 233–245 (2007).
4. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
5. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
6. Sawyers, C. L. Chronic myeloid leukemia. *N. Engl. J. Med.* **340,** 1330–1340 (1999).
7. Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and *ETS* transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
8. Soda, M. *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* **448,** 561–566 (2007).
9. Höglund, M., Gisselsson, D., Sall, T. & Mitelman, F. Coping with complexity. multivariate analysis of tumor karyotypes. *Cancer Genet. Cytogenet.* **135,** 103–109 (2002).
10. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17,** 1296–1303 (2007).
11. Volik, S. *et al.* End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA* **100,** 7696–7701 (2003).
12. Ruan, Y. *et al.* Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* **17,** 828–838 (2007).
13. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
14. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
15. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
16. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450,** 893–898 (2007).
17. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).

1009

18. Weterings, E. & Chen, D. J. The endless tale of non-homologous end-joining. *Cell Res.* **18**, 114–124 (2008).

19. van Gent, D. C. & van der Burg, M. Non-homologous end-joining, a sticky affair. *Oncogene* **26**, 7731–7740 (2007).

20. Hefferin, M. L. & Tomkinson, A. E. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair* **4**, 639–648 (2005).

21. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nature Rev. Genet.* **10**, 551–564 (2009).

22. Yan, C. T. *et al.* IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature* **449**, 478–482 (2007).

23. Bohlander, S. K. ETV6: a versatile player in leukemogenesis. *Semin. Cancer Biol.* **15**, 162–174 (2005).

24. Knezevich, S. R., McFadden, D. E., Tao, W., Lim, J. F. & Sorensen, P. H. A novel *ETV6–NTRK3* gene fusion in congenital fibrosarcoma. *Nature Genet.* **18**, 184–187 (1998).

25. Lannon, C. L. & Sorensen, P. H. ETV6–NTRK3: a chimeric protein tyrosine kinase with transformation activity in multiple cell lineages. *Semin. Cancer Biol.* **15**, 215–223 (2005).

26. Jones, D. T. *et al.* Tandem duplication producing a novel oncogenic *BRAF* fusion gene defines the majority of pilocytic astrocytomas. *Cancer Res.* **68**, 8673–8677 (2008).

27. Basecke, J., Whelan, J. T., Griesinger, F. & Bertrand, F. E. The MLL partial tandem duplication in acute myeloid leukaemia. *Br. J. Haematol.* **135**, 438–449 (2006).

28. Blow, J. J. & Gillespie, P. J. Replication licensing and cancer—a fatal entanglement? *Nature Rev. Cancer* **8**, 799–806 (2008).

29. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

30. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).

31. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).

32. Bergamaschi, A. *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosom. Cancer* **45**, 1033–1040 (2006).

33. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).

Author Contributions M.R.S., P.A.F., P.J.C. and P.J.S. designed the experiment. S.E., D.J.M., P.J.S., M-L.L., I.V., L.J.M., J.B., M.A.Q., H.S., C.C., R.N., A.M.S., A.L., J.W.M.M. and C.Latimer carried out laboratory analyses. J.A.F., J.S.R.-F., L.v.V., A.L.R. D.P.S. and A.-L.B.-D. provided clinical samples. P.J.S., D.J.M., I.V., M.-L.L., E.D.P., J.T.S., L.A.S., C.Leroy, C.D.G., M.J., J.W.T., K.W.L., P.J.C., P.A.F., J.S.R.-F., J.W.M.M., A.M.S., J.A.F., M.R.S., H.E.G.R., A.L.R., A.-L.B.-D., L.v.V., A.L., P.J.C. and P.A.F. performed data analysis, informatics and statistics. M.R.S. wrote the manuscript with comments from P.J.S., P.A.F., P.J.C., A.-L.B.-D., J.S.R.-F., J.A.F., A.L.R., D.P.S. and L.v.V.

METHODS

Library construction and paired-end sequencing. Genomic libraries from nine breast cancer cell lines and fifteen primary breast cancers were generated using 5 μg of total genomic DNA[17]. Briefly, 5 μg of genomic DNA was randomly fragmented to between 200 and 700 bp by focused acoustic shearing (Covaris Inc.). These fragments were electrophoresed on a 2% agarose gel where from the 450–550-bp fraction was excised and extracted using the Qiagen gel extraction kit (with gel dissolution in chaotropic buffer at room temperature to ensure recovery of (A+T)-rich sequences). The size-fractionated DNA was end repaired using T4 DNA polymerase, Klenow polymerase and T4 polynucleotide kinase. The resulting blunt-ended fragments were A-tailed using a 3′–5′ exonuclease-deficient Klenow fragment and ligated to Illumina paired-end adaptor oligonucleotides in a 'TA' ligation at room temperature for 15 min. The ligation mixture was electrophoresed on a 2% agarose gel and size-selected by removing a 2-mm horizontal slice of gel at ~600 bp using a sterile scalpel blade. DNA was extracted from the agarose as above. Ten nanograms of the resulting DNA was PCR-amplified for 18 cycles using 2 units of Phusion polymerase. PCR cleanup was performed using AMPure beads (Agencourt BioSciences Corporation) following the manufacturer's protocol. We prepared Genome Analyser paired-end flow cells on the supplied Illumina cluster station and generated 37-bp paired-end sequence reads on the Illumina Genome Analyser platform following the manufacturer's protocol. Images from the Genome Analyser were processed using the manufacturer's software to generate FASTQ sequence files. These were aligned to the human genome (NCBI build 36.2) using the MAQ algorithm v0.4.3[33].

Reads removed from structural variant analysis. Reads that failed to align in the expected orientation or distance apart were further evaluated using the SSAHA algorithm[34] to remove mapping errors in repetitive regions of the genome. In addition, during the PCR enrichment step, multiple PCR products derived from the same genomic template can occasionally be sequenced. To remove these, reads where both ends mapped to identical genomic locations (plus or minus a single nucleotide) were considered PCR duplicates, and only the read pair with the highest mapping quality retained. Further, erroneous mapping of reads originating from DNA present in sequence gaps in NCBI build 36.2 were removed by excluding the highly repetitive regions within 1 Mb of a centromeric or telomeric sequence gap. Additional read pairs, where both ends mapped to within less than 500 bp of one another, but in the incorrect orientation, were excluded from analysis, unless support for a putative rearrangement was indicated by additional read pairs. The majority of these singleton read pairs are likely to be artefacts resulting from either intramolecular rearrangements generated during library amplification or mispriming of the sequencing oligonucleotide within the bridge amplified cluster. Finally, read pairs where both ends mapped to within 500 bp of a previously identified germline structural variant were removed from further analysis, as these were likely to represent the same germline allele.

Generation of genome-wide copy number plots. Full methods for generation of high-resolution, genome-wide copy number information can be found in ref. 17. Briefly, the human reference genome was divided into bins of ~15 kb of mappable sequence and high-quality, correctly mapping read pairs, with a MAQ alternative mapping quality ≥35, were assigned to their correct bin and plotted. A binary circular segmentation algorithm originally developed for genomic hybridization microarray data[35] was applied to these raw plots to identify change points in copy number by iterative binary segmentation.

PCR confirmation of putative rearrangements. The following criteria were used to determine which incorrectly mapping reads pairs were evaluated by confirmatory PCR: (1) reads mapping ≥10 kb apart spanned by ≥2 read-independent read pairs (where at least one read pair had an alternative mapping quality ≥35); (2) reads mapping ≥10 kb apart spanned by 1 read pair (with an alternative mapping quality ≥35), with both ends mapping to within 100 kb of a change point in copy number identified by the segmentation algorithm; (3) reads mapping ≥600 bp apart spanned by ≥2 read-independent read pairs (where at least one read pair had an alternative mapping quality ≥35) with both ends mapping to within 100 kb of a change point in copy number identified by the segmentation algorithm; (4) selected read pairs mapping between 600 bp and 10 kb apart spanned by ≥2 independent read pairs (where at least one read pair had an alternative mapping quality ≥35).

Primers were designed to span the possible breakpoint and to generate a maximum product size of 1 kb. PCR reactions were performed on tumour and normal genomic DNA for each set of primers at least twice, using the following thermocycling parameters: 95 °C × 15 min (95 °C × 30 s, 60 °C × 30 s, 72 °C × 30 s) for 30 cycles, 72 °C × 10 min. Products giving a band were sequenced by conventional Sanger capillary methods and compared to the reference sequence to identify breakpoints. Somatically acquired rearrangements were defined as those generating a reproducible band in the tumour DNA with no band in the normal DNA following PCR amplification, together with unambiguously mapping sequence data suggesting a rearrangement.

RT–PCR and cloning. Total RNA (100 ng) from the tumour and matched constitutional DNA/lymphoblastoid cell lines was reverse transcribed into single-stranded cDNAs using Reverse Transcriptase II (Invitrogen) and Oligo (dT)$_{12–18}$ (Invitrogen) in 20 μl reaction at 25 °C for 10 min, 42 °C for 50 min, 72 °C for 15 min. The cDNA was then diluted with 30 μl of distilled water before subsequent PCR amplification. Resulting bands were sequenced to confirm the specificity of the reaction and the presence of the aberrant transcript. To detect fusion transcripts, we used forward primers in the putative 5′ partner gene and reverse primers from the 3′ partner. To detect rearranged transcripts, we used forward primers and reverse primers corresponding to the predicted exons fused. When multiple bands, possibly suggestive of splice variants, were detected, all bands were excised from the gel and capillary sequenced separately.

DNA probes and FISH. FISH based on the split apart or fusion probe strategy was used to validate *NFIA–EHF* gene aberrations. For the *NFIA–EHF* fusion probe, the following BAC clones (http://www.ensembl.org/, Ensembl release 54) were used: RP11-32I17, chromosome 1 (61,191,261–61,339,873), RP11-364M11, chromosome 1 (61,064,196–61,228,554) (red); and RP11-64P01, chromosome 1 (34,699,699–34,860,527), RP11-277N08, chromosome 1 (34,722,104–34,965,946) (green). For the *EHF* split-apart probe, the following BAC clones were used: RP11-64P01, chromosome 11 (34,699,699–34,860,527), RP11-277N08, chromosome 11 (34,772,104–34,965,946) (green); and RP11-567H10, chromosome 11 (34,123,423–34,294,895), RP11-278N12, chromosome 11 (34,086,610–34,248,310) and RP11-686L07, chromosome 11 (33,936,895–34,109,642) (red).

Dual colour FISH was used to detect the *SLC26A6–PRKAR2A* tandem duplication using BAC clones RP11-527M10, chromosome 3 (45,948,424–46,115,480) (green) and RP11-148G20, chromosome 3 (48,575,991–48,781,362) (red).

BAC clones were purchased from BACPAC resource (Children's Hospital Oakland; http://bacpac.chori.org/), and DNA from all BAC clones was purified, labelled and individually verified for specificity by FISH and direct sequencing as described previously[36]. BAC DNA was labelled with either biotin or digoxygen-lI-dUTP (Roche) using the Bioprime kit (Invitrogen) and FISH was performed as described previously[36]. Biotinylated probes were detected with Cy5–Streptavidin (Invitrogen, Zymed Laboratories) and digoxigenin-labelled BACs, with anti-digoxigenin-fluorescein, Roche (green). Nuclei and chromosomes were counterstained with DAPI. Images were captured with a Zeiss Axioplan 2 microscope equipped with a CCD camera (Applied Imaging Diagnostic Instruments) and Cytovision software, version 2.81 (Applied Imaging). Only morphologically intact and non-overlapping nuclei were analysed.

Breakpoint analysis. All 1,832 breakpoints defined to the base-pair level were used in the analysis of breakpoint sequence context, excluding shards and overlapping regions. Analysis was performed on all breakpoints together, and also on subsets divided into deletions, tandem duplications, amplicons, other intrachromosomal events, and all interchromosomal events. 10 bp and 100 bp on either side of the breakpoint sites were extracted for analysis. As a control, for each real breakpoint, 100 sequences of the same length were extracted from the regions extending from 10,000 to 20,000 bp away from either side of the break. These matched control sequences were used as a comparison in the analysis to account for any regional differences such as large variations in GC or repeat content. The length of nucleotide tracts (polynucleotide, polypurine/polypyrimidine, and alternating polypyrimidine/polypurine) was compared in the breakpoint and control regions using a one-tailed Mann–Whitney U-test, and the average GC content and presence of known motifs associated with DNA breaks[37] were compared using a Fisher exact test.

Enrichment of breakpoints in genes. To determine whether breakpoints were enriched in genic regions, we compared the number of breakpoints falling within genes to an empirically derived expected proportion. We classified breakpoints as genic or intergenic based on if their coordinates fell within a gene as annotated by Ensembl (http://www.ensembl.org/, Ensembl release 54). To account for the fact that some areas of the genome will be difficult to sequence align to with short reads, we derived the expected proportion of breakpoints that should fall within a gene from the actual proportion of read pairs that aligned to genic regions. Treating each breakpoint of a rearrangement independently, we then compared the number of breakpoints falling within a gene to this expected proportion using a Chi-squared test to obtain a P-value for the overrepresentation of breakpoints in genes.

Inter-individual heterogeneity in the patterns of microhomology. Comparison of microhomology and non-templated sequence distributions across individual samples was performed using Scholz and Stephen' k-sample generalization of the Anderson–Darling goodness-of-fit test, with 10,000 data permutations to generate the statistic's null distribution[38].

34. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).

35. Venkatraman, E. & Olshen, A. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

36. Lambros, M. *et al.* Unlocking pathology archives for molecular genetic studies: a reliable method to generate probes for chromogenic and fluorescent *in situ* hybridization. *Lab. Invest.* **86**, 398–408 (2006).

37. Abeysinghe, S. S. *et al.* Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum. Mutat.* **22**, 229–244 (2003).

38. Scholz, F. W. & Stephens, M. A. K-sample Anderson-Darling tests. *Am. J. Stat. Assoc.* **82**, 918–924 (1987).

ARTICLES

Structure of the outer membrane complex of a type IV secretion system

Vidya Chandran[1]*, Rémi Fronzes[1]*, Stéphane Duquerroy[2,3], Nora Cronin[1], Jorge Navaza[4] & Gabriel Waksman[1]

Type IV secretion systems are secretion nanomachines spanning the two membranes of Gram-negative bacteria. Three proteins, VirB7, VirB9 and VirB10, assemble into a 1.05 megadalton (MDa) core spanning the inner and outer membranes. This core consists of 14 copies of each of the proteins and forms two layers, the I and O layers, inserting in the inner and outer membrane, respectively. Here we present the crystal structure of a ~0.6 MDa outer-membrane complex containing the entire O layer. This structure is the largest determined for an outer-membrane channel and is unprecedented in being composed of three proteins. Unexpectedly, this structure identifies VirB10 as the outer-membrane channel with a unique hydrophobic double-helical transmembrane region. This structure establishes VirB10 as the only known protein crossing both membranes of Gram-negative bacteria. Comparison of the cryo-electron microscopy (cryo-EM) and crystallographic structures points to conformational changes regulating channel opening and closing.

Type IV secretion (T4S) systems are used by Gram-negative bacteria in a variety of processes, ranging from the delivery of virulence factors into eukaryotic cells to conjugative transfer of genetic material and the uptake or release of DNA[1–3]. In *Helicobacter pylori*, *Brucella suis* and *Legionella pneumophila*, T4S systems mediate the injection of virulence proteins into mammalian host cells to cause gastric ulcers, brucellosis, or Legionnaire's disease, respectively[4,5]. In *Agrobacterium tumefaciens*, the VirB/D T4S system delivers oncogenic DNA and proteins into plant cells[2,6], and *Bordetella pertussis* uses a T4S system to secrete the pertussis toxin into the extracellular milieu[7]. Conjugation promotes bacterial genome plasticity and the adaptive response of bacteria to changes in the environment, contributing to the spread of antibiotic-resistance genes among pathogenic bacteria[8].

Although variations exist, many of the T4S systems found in Gram-negative bacteria are similar to the *A. tumefaciens* VirB/D T4S system, which comprises 12 proteins named VirB1 to VirB11 and VirD4[2]. In general (but not always), T4S systems include an extracellular pilus composed of a major (VirB2) and a minor (VirB5) subunit[9]. Three ATPases located at the inner membrane, VirB4, VirB11, and VirD4, power substrate secretion and possibly assist in the assembly of the system[2,10]. The inner-membrane channel is thought to be composed of the polytopic membrane protein VirB6 and the bitopic membrane proteins VirB8 and VirB10[2]. At the outer membrane, the composition of the pore that allows the substrate to reach the extracellular milieu is unknown. VirB9 in complex with the short lipoprotein VirB7 could be part of this structure. A region of the carboxy (C)-terminal domain of VirB9 was shown to be surface-exposed[11]. However, no transmembrane region could be found or predicted in both proteins.

Recently, the cryo-EM structure of the core complex of the T4S system encoded by the conjugative plasmid pKM101 showed that T4S systems consist of a 1.05 MDa core spanning the inner and outer membranes of Gram-negative bacteria[12]. This core, extracted pre-assembled from the membranes using detergents, is composed of two layers, the O and I layers, and is formed by 14 copies of three proteins, the homologues of the VirB7, VirB9 and VirB10 proteins,

termed TraN, TraO and TraF, respectively. Here we present the crystal structure of the T4S system outer-membrane complex, containing the entire O layer.

General architecture of the complex

The T4S system outer-membrane complex was obtained from chymotryptic cleavage of the T4S system core, crystallized and its structure solved as described in the Methods (see also Supplementary Table 1 and Supplementary Fig. 1). This 590 kDa complex contains 14 copies each of the C-terminal domain of TraF (TraF$_{CT}$; residues 160–386), the C-terminal domain of TraO (TraO$_{CT}$; residues 160–294) and the full-length TraN (Supplementary Fig. 1a). The hetero-tetradecameric structure of the T4S system outer-membrane complex has an elegant architecture with 14-fold symmetry (Fig. 1a). In the cryo-EM work, the O layer was described as containing two parts: the main body and the cap. In the crystal structure, the cap is made of a hydrophobic ring of two-helix bundles defining a 32 Å channel (Fig. 1a, b). These helices are made of residues 307–355 of TraF$_{CT}$. The main body is made of the rest of TraF$_{CT}$, TraO$_{CT}$ and TraN. It has a 172 Å diameter (Fig. 1b). When the complex is viewed from the top, that is, from the extracellular milieu (Supplementary Fig. 2a), TraF$_{CT}$ forms an inner ring surrounded by the TraO$_{CT}$–TraN complex. The TraN protein forms spokes radially crossing the entire assembly (Supplementary Fig. 2a, b). When the complex is viewed from the bottom, that is, from the periplasm (Supplementary Fig. 2c), the structure is primarily made of TraF$_{CT}$. A cut-away view of the structure (Fig. 1b) shows that TraF forms the entirety of the inner wall of the structure, establishing VirB10 homologues as T4S system outer-membrane channel proteins, an unexpected result as VirB10 homologues have never been suspected to form the outer-membrane channel of T4S systems. VirB10 homologues are also found in the inner-membrane fraction owing to the presence of a transmembrane helix at the amino (N) terminus of the protein[13]. Thus, VirB10 is the only protein known to insert in both inner and outer membranes of Gram-negative bacteria.

[1]Institute of Structural and Molecular Biology, University College London and Birkbeck College, Malet Street, London WC1E 7HX, UK. [2]Institut Pasteur, Unité de Virologie Structurale, Virology Department and CNRS URA 3015, Paris, 25–28 Rue du Dr Roux, F-75724 Paris, France. [3]Université Paris-Sud, F-91405 Orsay, France. [4]Laboratoire de Microscopie Electronique, Institut de Biologie Structurale J.P. Ebel, 41 rue Jules Horowitz, F-38027 Grenoble Cedex 1, France.
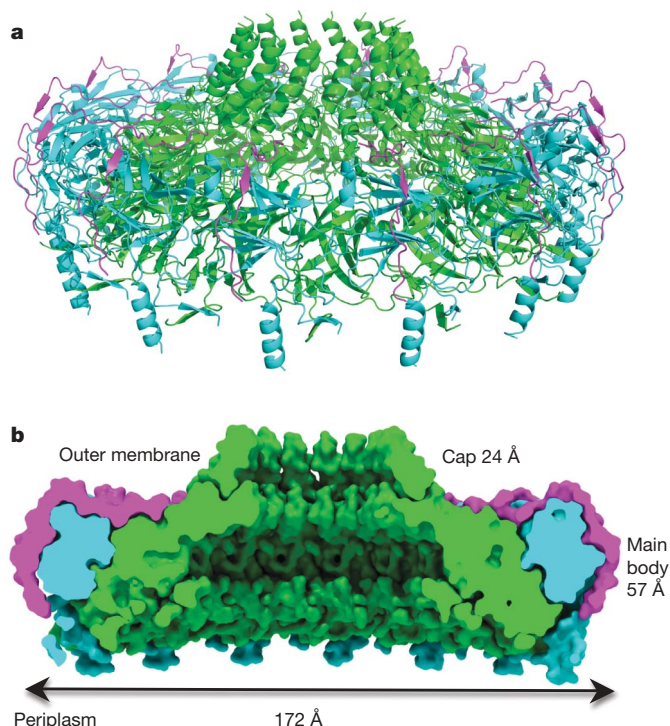*These authors contributed equally to this work.

Figure 1 | The T4S system outer-membrane complex. Ribbon diagram (**a**) and space-filling cut-away (**b**) of the tetradecameric complex. TraF$_{CT}$, TraO$_{CT}$ and TraN subunits are colour-coded green, cyan and magenta, respectively. In **b**, dimensions and labelling of the various parts of the complex are provided.

Structure of the heterotrimer unit

A unique feature of the T4S system outer-membrane complex is that it is made of three proteins (instead of one for all outer-membrane structures previously determined), all three being indispensable for complex assembly and channel formation[12]. The heterotrimer formed by TraF$_{CT}$, TraO$_{CT}$ and TraN is shown in Fig. 2 (see also Supplementary Fig. 2d). The interfaces between TraF$_{CT}$ and TraO$_{CT}$ and between TraO$_{CT}$ and TraN bury 2861 Å^2 and 2024 Å^2 of surface area, respectively. TraN does not interact with TraF$_{CT}$ in the heterotrimer. The core structure of the TraO$_{CT}$–TraN complex bound to TraF$_{CT}$ is similar to the previously determined NMR structure of the TraO$_{CT}$–TraN complex alone (Fig. 2; details in Supplementary Information and Supplementary Fig. 3a, b)[11]. However, there are also important differences. First, in TraO$_{CT}$ bound to TraF$_{CT}$, residues 162–173 at the N terminus form the α1 helix (Fig. 2). In the tetradecamer, 14 α1 helices form 14 pillars onto which the entire complex sits on the I layer (Fig. 1a). Second, at the C terminus, residues 286–289 have become ordered in TraO$_{CT}$ bound to TraF$_{CT}$ to form the β10 strand. β10 of TraO$_{CT}$ is part of the TraF$_{CT}$–TraO$_{CT}$ interface as it forms a two-stranded β-sheet with β8 of TraF$_{CT}$ (see Fig. 2 where, for clarity, only β8 of TraF$_{CT}$ is indicated). In TraN, sequences at the N (residues 15–25) and C (residues 37–42) termini have become ordered in the heterotrimer. Remarkably, TraN almost entirely wraps around TraO$_{CT}$, forming an interface with TraO$_{CT}$ that is much more extensive than that observed in the NMR structure (see details in Supplementary Information and Supplementary Fig. 3c).

The structure of the apo form of the C-terminal domain of ComB10 (ComB10$_{CT}$), a TraF/VirB10 homologue, has been solved[14]. It consists of an atypical β-barrel flanked by an α-helix, onto which an 'antenna'-like structure is mounted (Supplementary Fig. 3d). TraF$_{CT}$ in the heterotrimer differs from ComB10$_{CT}$ in three major ways: (1) the flanking helix of ComB10$_{CT}$ is missing in TraF$_{CT}$; (2) the antenna structure of TraF$_{CT}$ is more extended; and (3) the N terminus of TraF$_{CT}$ forms an extended N-terminal arm (or lever arm; residues

Figure 2 | Ribbon diagram of the heterotrimer unit. Structures mentioned in the main text are indicated. The insert locates the shown heterotrimer within the tetradecameric structure. A stereo version of this figure with full secondary structure labelling is provided in Supplementary Fig. 2d.

171–199) that projects out to interact with three consecutive neighbouring heterotrimers in the tetradecameric assembly (Figs 2 and 3, and Supplementary Figs 3e and 4).

Tetradecameric assembly

The interactions between heterotrimers in the tetradecameric outer-membrane complex are extensive (details are provided in Supplementary Figs 4–7 and Supplementary Information). To help orientate the reader, a consistent colour scheme summarized in Fig. 3a, e is provided, in which the TraF/VirB10, TraO/VirB9 and TraN/VirB7 subunits are numbered clockwise F1–F14, O1–O14 and N1–N14, respectively.

Of the 13,000 Å^2 of total surface area of TraF$_{CT}$, 11% is involved in intra-heterotrimeric interactions and 51.5% is involved in interactions between heterotrimers (Supplementary Fig. 8a, b). TraF$_{CT}$ of one heterotrimer (for example F1 in green in Fig. 3a, b) interacts not only with two adjacent TraF$_{CT}$ subunits (F2 and F14), but also, because of the long N-terminal lever arm, with four neighbouring TraF$_{CT}$ subunits located further afield (F3, F4 on one side and F13 and F12 on the other; Fig. 3b). In addition, TraF$_{CT}$ interacts with the TraO$_{CT}$ and TraN subunits of an adjacent heterotrimer (O14 and N14 in Fig. 3a)

The interface between TraF$_{CT}$ subunits includes 2320 Å^2 of surface area and is described in detail in the Supplementary Information and in Supplementary Fig. 6. Strand additions and numerous loop–loop interactions constitute important parts of the interface. Yet, its most notable feature is the N-terminal lever arm of the TraF$_{CT}$ subunit, which makes its own extensive interaction network. The arm of subunit FX ($1 \le X \le 14$) interacts with the subunit F$X + 1$, F$X + 2$ and F$X + 3$ (for example, see Fig. 3b, c where the F1 lever arm interacts with subunits F2, F3 and F4). FX − F$X + 1$, FX − F$X + 2$ and FX − F$X + 3$ contacts involve residues in the β$_n$1 strand, the β$_n$1 − α$_n$1 linker, and α$_n$1 helix of the FX subunit lever arm, respectively (Fig. 3c and Supplementary Fig. 6c, d). Overall, the N-terminal lever arms form a continuous inner shelf at the base of the outer-membrane complex (Fig. 3d).
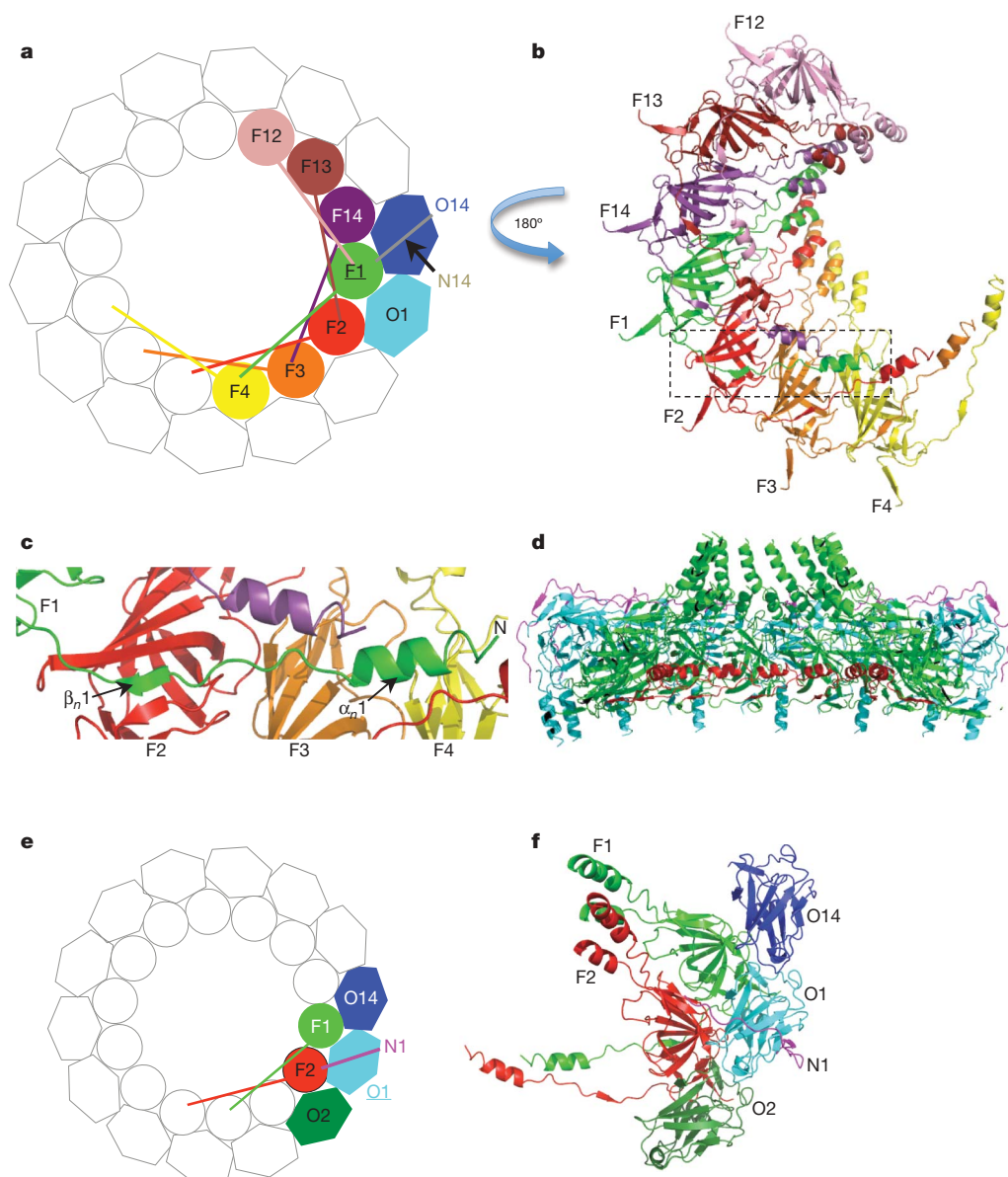
Figure 3 | Inter-heterotrimeric interactions. a, Schematic diagram of the tetradecamer with emphasis on interactions with the $TraF_{CT}$ subunit in green (F1). $TraF_{CT}$ and $TraO_{CT}$ subunits are shown in circles and hexagons, respectively. Not shown for clarity: TraN subunits (except N14, which makes interactions with F1) and the lever arms of subunits not interacting with F1. **b**, Ribbon diagram of the F1-interacting $TraF_{CT}$ subunits (that is, F4, F3, F2, F14, F13 and F12) viewed from the periplasm; that is, turned 180° compared with the view in **a**. The rectangle locates the lever arm of subunit F1. **c**, Interactions of subunits F2, F3 and F4 with the lever arm of subunit F1. The secondary structures of the lever arm of F1 are labelled. The F2, F3 and F4 subunits are in ribbon representation and labelled accordingly. **d**, Cut-away side view of the outer-membrane complex with the proteins in ribbon diagram representation colour-coded as in Fig. 1 but for the N-terminal arms of the $TraF_{CT}$ subunits in red. **e**, Schematic diagram of the tetradecamer with emphasis on interactions with the $TraO_{CT}$ subunit in cyan (O1). Not shown for clarity: TraN subunits (except N1, which makes interactions with O1) and the lever arms of subunits not interacting with O1. **f**, Ribbon diagram of the subunits shown in colour at left. This view is from the extracellular milieu.

In the tetradecameric structure of the outer-membrane complex, $TraO_{CT}$ interacts with five proteins: two adjacent $TraO_{CT}$ subunits, 1 $TraF_{CT}$ subunit of an adjacent heterotrimer, 1 $TraF_{CT}$ and 1 TraN within its own heterotrimer (Fig. 3e, f). Overall, 49.25% of the total surface area of $TraO_{CT}$ is involved in protein–protein interaction, 43% of which is with the adjacent heterotrimer proteins (Supplementary Fig. 8c, d). The interfaces between adjacent $TraO_{CT}$ subunits and between $TraO_{CT}$ of one heterotrimer and $TraF_{CT}$ of an adjacent heterotrimer (described in the Supplementary Information and Supplementary Fig. 7) mostly consist of loop residues (Fig. 3f).

The outer-membrane pore

Viewed from the top, the structure contains a central hydrophobic ring of 76 Å in diameter (Fig. 4a, left panel) with a 32 Å pore in the middle. This region consists of a ring of two-helix bundles (Fig. 4a, lower right panel) formed by helices α2 and α3 of the $TraF_{CT}$ antenna. These helices each form a ring, with the α3 ring inside the α2 ring, suggesting that α2 is the transmembrane helix contacting the membrane. Two lines of evidence indicate that this region is inserted in the outer membrane and forms the outer-membrane channel. First, α2, the external helix in the two-helix bundle, has all the features of a transmembrane helix: it is amphipathic, with its hydrophobic side expected to contact the membrane, and is strongly predicted to form a transmembrane helix by TMPred (http://www.ch.embnet.org/software/TMPRED_form.html). Second, when a Flag-tag is inserted between helices α2 and α3 of TraF in the full-length core complex, the tag is found exposed extracellularly (Fig. 4b), demonstrating that the two-helix bundle projects across the outer membrane. Altogether, these results unambiguously assign the two-helix bundle region of TraF/VirB10 as the channel-forming region of T4S systems.
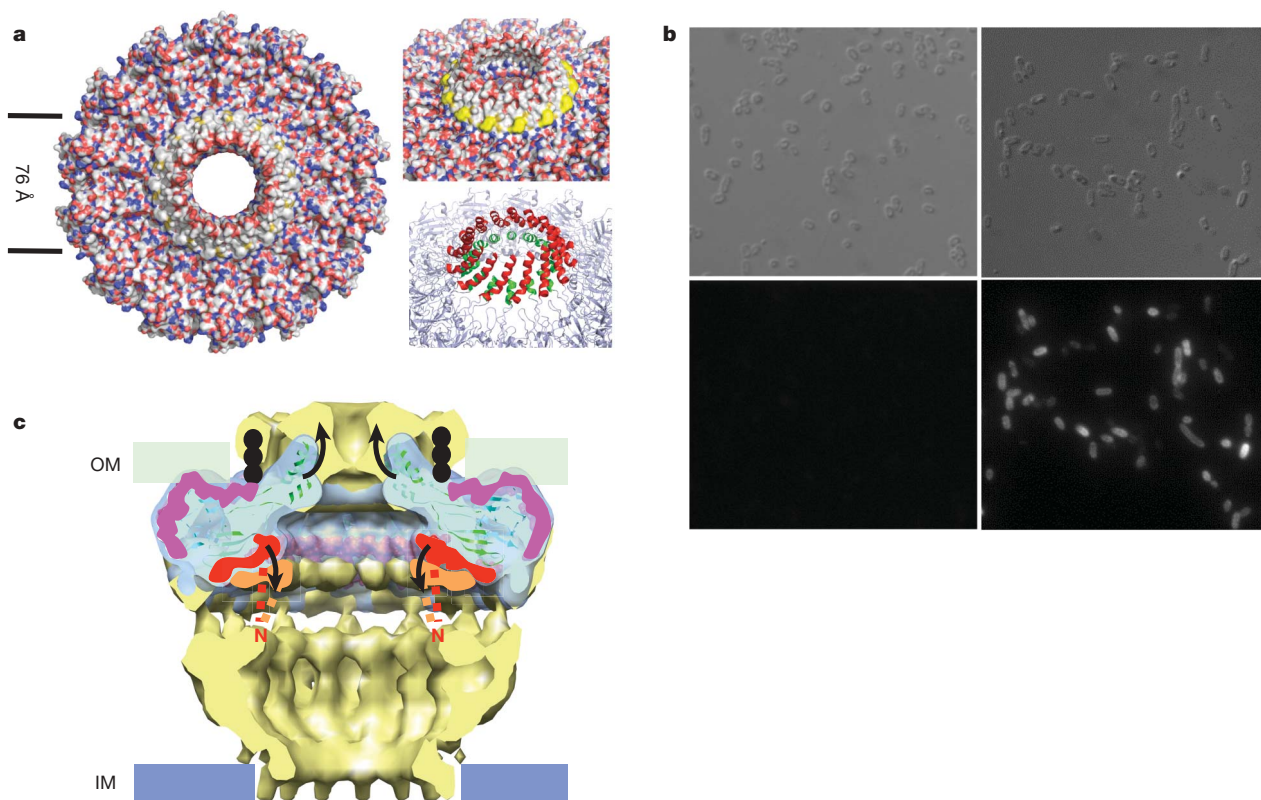
1013

Figure 4 | Transmembrane region of the T4S system outer-membrane complex and proposed mechanism of pore opening and closure. a, Left panel: surface diagram of complex viewed from the extracellular milieu. The two lines define the hydrophobic central region around the pore. Top right panel: same as at left but showing (in yellow) the ring of Trp residues at the base of α2. Bottom right panel: ribbon diagram of the transmembrane helices with the internal and external ring of helices colour-coded in green and red, respectively. **b**, Extracellular detection of the Flag-tag located between helices α2 and α3 of TraF. A Flag-tag was introduced in the loop between the α2 and α3 helices of TraF as described in Methods. Upper and lower left panels: control sample expressing the wild-type T4S system core complex. Upper and lower right panels: cells expressing the Flag-tagged T4S system core complex. Differential interference contrast images are shown in the upper panels and the corresponding fluorescence images in the lower panels. **c**, Proposed mechanism for pore opening and closure. The cryo-EM core complex structure (pale yellow) is superimposed on the crystal structure of the outer-membrane complex (pale cyan). The N-terminal lever arm shelf is highlighted in orange and red in the cryo-EM and crystal structures, respectively. The position of the lipid on TraN/VirB7 is indicated in black dots. TraN/VirB7 is shown in magenta. Proposed conformational changes are illustrated by arrows. The N-terminal sequences linking the lever arms to sequences in the I layer are shown in dashed lines colour-coded red and orange for the conformations found in the crystal and cryo-EM structures, respectively. OM, outer membrane; IM, inner membrane.

α-Helical insertions into outer membranes have been observed only once before, in Wza[15,16]. In Wza, the C terminus forms an α-helix and eight of them in the Wza octamer were shown unambiguously to form the outer-membrane channel. The outer-membrane channel of the T4S system differs by forming a two-helix bundle ring system. As in Wza, a tryptophan residue lies at the N-terminal base of the TraF transmembrane helix contacting the membrane (α2; Fig. 4a, upper right panel)[15]. This Trp residue is conserved in most VirB10 protein family members. In the structure presented here, the outer-membrane-inserting region has been cleaved by chymotrypsin (Fig. 1a) and the region around the cleavage is disordered (residues 322–344 could not be traced because of poor electron density). As a result, the transmembrane α2 helices are one turn shorter than those forming the outer-membrane channel of Wza. Another consequence of this structural disorder is that the extent of the opening cannot be precisely defined.

The T4S system outer-membrane complex, in addition to being much larger than previously described outer-membrane structures (for example Wza or TolC), is also radically different in shape (Supplementary Fig. 9): whereas the Wza homo-octamer and the homo-TolC trimer are elongated[15,17], the hetero-tetradecameric T4S system outer-membrane complex spreads out just under the outer membrane, creating a vast contact area with the inner leaflet of the outer membrane. This contact area is mostly mediated by the TraO$_{CT}$–TraN complex ring. In effect, the TraO$_{CT}$–TraN complex ring appears to buttress the entire T4S system against the inner leaflet of the outer membrane, possibly allowing the system to exert much greater mechanical forces to extrude substrates than would otherwise be feasible in systems like Wza or TolC, which make limited contacts with the inner leaflet of the outer membrane.

Conformational changes in T4S systems

The cryo-EM structure of the T4S system core complex revealed a double-walled structure in the cap of the O layer (Supplementary Fig. 10a)[12]. The crystal structure of the cap is not double-walled. This is because the lipid moiety of TraN/VirB7 was not built (the single-wavelength anomalous dispersion (SAD)-derived map showed electron density connecting to Cys 15 of TraN/VirB7 (the lipidation site) but the density was not sufficiently well-defined for unambiguous fitting). However, when the crystal structure of the outer-membrane complex is superimposed on the cryo-EM structure of the entire complex (Fig. 4c), Cys 15 of TraN aligns perfectly at the base of the outer wall of the cap, suggesting that this part of the cap might indeed contain the lipidated part of TraN (indicated in Fig. 4c in black dots).

The superposition of the structure of the outer-membrane complex and that of the core complex (Fig. 4c; such a superposition is valid as one complex is derived from the other by proteolysis) also reveals that the transmembrane helices overlap partly with the first half of the inner wall of the cap, suggesting that they might form this region of the cap. However, the crystal structure captures them in a different conformation than that observed in the cryo-EM structure: whereas in the cryo-EM structure they form a narrow vertical constriction, in the crystal structure they lie in a 'relaxed' state at a 45°

angle. Thus, the two structures may represent different states. Presumably, by removing the entire N-terminal half of TraF, the constraints that this half places on the N-terminal lever arm of $TraF_{CT}$ might have been removed, releasing the lever arms and leading to relaxation of the transmembrane helices. Indeed, as illustrated in Fig. 4c, sequences N-terminal to the $\alpha_n 1$ helix would directly connect to the I layer (schematically shown in Fig. 4c by dashed red lines) and are likely to bring the N-terminal arm of each TraF/VirB10 subunit down. This is consistent with the fact that the shelf formed by the N-terminal lever arms of TraF/VirB10 subunits has shifted up in the crystal structure compared with its position in the cryo-EM structure of the full-length core complex (Fig. 4c and Supplementary Fig. 10a, b in orange and red, respectively). Thus, we propose that the N-terminal arms of the TraF/VirB10 subunits might act in concert to exert conformational changes in the T4S system channel upon signals sensed by the N-terminal domain of the protein (see below).

In *A. tumefaciens*, VirB10 is known to undergo a conformational change induced by the energizing T4S system components[18]. The structure of the T4S system outer-membrane complex reveals that VirB10 forms the outer-membrane channel. However, VirB10 is also known to insert in the inner membrane, making contact not only with the inner-membrane channel component VirB8, but also with the ATPases[13,19,20]. Thus, VirB10 is in a unique position to relay conformational changes taking place in the ATPases and to effect pore opening and closure at the outer membrane. Also, in *A. tumefaciens*, VirB10 does not directly contact the substrate, but regulates its handover from the VirB6/VirB8 complex in the inner membrane to VirB9 and VirB2, the major pilin[21]. As VirB10 lines the interior of the outer-membrane complex, it is difficult to envisage how the substrate could not interact with it, unless it is insulated from the substrate by another layer of protein, presumably made of the VirB2 pilin. We thus propose that the VirB2 pilin forms a cylindrical conduit encased within the VirB10 ring. This hypothesis is also consistent with the observation that, in the state captured in the crystal structure (where VirB2 is absent), the TraF/VirB10 transmembrane helices have somewhat caved in.

The crystal structure of the T4S system outer-membrane complex reveals an outer-membrane structure of unprecedented size and complexity. This structure is held together by a dense network of protein–protein interactions, which provides a rich targeting ground for inhibitor design. Most striking among them is the extensive interaction that the N-terminal lever arms of the $TraF_{CT}$ subunits make with numerous subunits along the tetradecameric structure. We hypothesized that these sequences are at the heart of a nanodevice regulating T4S. If confirmed, this mechanism could become key to the design of inhibitor compounds specifically targeting the T4S machinery.

METHODS SUMMARY

Purification, crystallization, X-ray diffraction data collection. The T4S system core complex was expressed and purified as described previously[12]. After addition of chymotrypsin, the T4S system outer-membrane complex was purified by gel filtration. Crystals were grown by hanging-drop vapour diffusion. Native and SAD data at the selenium edge were collected at the European Synchrotron Radiation Facility beamline ID14.4 and Soleil's beamline PROXIMA 1, respectively, and processed using the XDS package.
Structure determination and refinement. The structure was solved by molecular replacement using the native data set and the known 20 Å resolution cryo-EM map of the trypsin-cleaved core complex as search model[12]. Fourteen-fold non-crystallographic symmetry averaging and phase extension to 2.8 Å resolution yielded a readily interpretable electron density map. This map was used to locate selenium atoms using the SAD data set. These heavy-metal sites were used to generate SAD phases to 2.6 Å resolution, also yielding an interpretable electron density map.
Extracellular localization of a Flag-tag inserted between the α2 and α3 helices of TraF. A Flag-tag was introduced at position 332 in the loop between the α2 and α3 helices of TraF. Cells expressing either the wild-type core complex or the Flag-tagged core complex were grown, fixed and incubated with anti-Flag antibodies followed by goat anti-mouse IgG1 Texas Red antibodies. Fluorescence

was monitored using a Zeiss Axioskop microscope, and images collected using a Hamamatsu Orca ER camera.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Fronzes, R., Christie, P. J. & Waksman, G. The structural biology of type IV secretion systems. *Nature Rev. Microbiol.* **7**, 703–714 (2009).
2. Christie, P. J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S. & Cascales, E. Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu. Rev. Microbiol.* **59**, 451–485 (2005).
3. Schroder, G. & Lanka, E. The mating pair formation system of conjugative plasmids – a versatile secretion machinery for transfer of proteins and DNA. *Plasmid* **54**, 1–25 (2005).
4. Backert, S. & Selbach, M. Role of type IV secretion in *Helicobacter pylori* pathogenesis. *Cell. Microbiol.* **10**, 1573–1581 (2008).
5. Ninio, S. & Roy, C. R. Effector proteins translocated by *Legionella pneumophila*: strength in numbers. *Trends Microbiol.* **15**, 372–380 (2007).
6. McCullen, C. A. & Binns, A. N. *Agrobacterium tumefaciens* and plant cell interactions and activities required for interkingdom macromolecular transfer. *Annu. Rev. Cell Dev. Biol.* **22**, 101–127 (2006).
7. Burns, D. L. Type IV transporters of pathogenic bacteria. *Curr. Opin. Microbiol.* **6**, 29–34 (2003).
8. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Rev. Microbiol.* **3**, 711–721 (2005).
9. Fronzes, R., Remaut, H. & Waksman, G. Architectures and biogenesis of non-flagellar protein appendages in Gram-negative bacteria. *EMBO J.* **27**, 2271–2280 (2008).
10. Gomis-Ruth, F. X. & Coll, M. Cut and move: protein machinery for DNA processing in bacterial conjugation. *Curr. Opin. Struct. Biol.* **16**, 744–752 (2006).
11. Bayliss, R. *et al.* NMR structure of a complex between the VirB9/VirB7 interaction domains of the pKM101 type IV secretion system. *Proc. Natl Acad. Sci. USA* **104**, 1673–1678 (2007).
12. Fronzes, R. *et al.* Structure of a type IV secretion system core complex. *Science* **323**, 266–268 (2009).
13. Jakubowski, S. J. *et al. Agrobacterium* VirB10 domain requirements for type IV secretion and T pilus biogenesis. *Mol. Microbiol.* **71**, 779–794 (2009).
14. Terradot, L. *et al.* Structures of two core subunits of the bacterial type IV secretion system, VirB8 from *Brucella suis* and ComB10 from *Helicobacter pylori. Proc. Natl Acad. Sci. USA* **102**, 4596–4601 (2005).
15. Dong, C. *et al.* Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein. *Nature* **444**, 226–229 (2006).
16. Meng, G., Fronzes, R., Chandran, V., Remaut, H. & Waksman, G. Protein oligomerization in the bacterial outer membrane. *Mol. Membr. Biol.* **26**, 136–145 (2009).
17. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B. & Hughes, C. Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **405**, 914–919 (2000).
18. Cascales, E. & Christie, P. J. *Agrobacterium* VirB10, an ATP energy sensor required for type IV secretion. *Proc. Natl Acad. Sci. USA* **101**, 17228–17233 (2004).
19. Atmakuri, K., Cascales, E. & Christie, P. J. Energetic components VirD4, VirB11 and VirB4 mediate early DNA transfer reactions required for bacterial type IV secretion. *Mol. Microbiol.* **54**, 1199–1211 (2004).
20. Llosa, M., Zunzunegui, S. & de la Cruz, F. Conjugative coupling proteins interact with cognate and heterologous VirB10-like proteins while exhibiting specificity for cognate relaxosomes. *Proc. Natl Acad. Sci. USA* **100**, 10465–10470 (2003).
21. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a DNA substrate. *Science* **304**, 1170–1173 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions V.C. produced the complex, optimized crystals, and built, refined and analysed the structure. R.F. designed the purification protocol, produced the complex, grew the first crystals, optimized crystals and analysed the structure. S.D. and J.N. solved the structure by molecular replacement and provided the electron density map. N.C. collected crystallographic data. G.W. supervised the work, analysed the structure and wrote the paper.

Author Information Structure factors and coordinates are deposited in the Protein Data Bank under accession number 3JQO. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.W. (g.waksman@ucl.ac.uk or g.waksman@bbk.ac.uk).

nature

METHODS

Purification of the outer-membrane complex. Expression of the T4S system core complex from the IBA3c:*traN-traF$_{C-ST}$* plasmid was induced using 200 μg l^{-1} of anhydrotetracyclin[12]. After incubation overnight at 16 °C, the complex was extracted from the membrane fraction and purified using a Strep-Tactin sepharose affinity column (IBA) as described previously[12]. The elution fractions containing the core complex were pooled together and subjected to limited proteolysis with 2 mg ml^{-1} of chymotrypsin for 3 h at room temperature. The proteolysed sample was then concentrated using a 100-kDa cut-off spin concentrator (Amicon) and loaded onto a Superose 6 GL 10/300 (GE Healthcare) gel filtration column in 50 mM TrisHCl pH 8.0, 200 mM NaCl and 10 mM LDAO. The outer-membrane complex eluted as a single peak (Supplementary Fig. 1a).

Selenomethionine labelling. The IBA3c:*traN-traF$_{C-ST}$* plasmid[12] was transformed into B834(DE3) competent cells (Novagen). Expression of the selenomethionine (SeMet)-labelled proteins was performed at 16 °C overnight in M9 minimal medium supplemented with 50 mg l^{-1} of SeMet. Purification of the SeMet-labelled outer-membrane complex was as described above for the native complex.

Crystallization and data collection. Large rod-like crystals were grown by vapour diffusion method at 20 °C using hanging drops containing 15 mg ml^{-1} of the purified native and SeMet-labelled complexes and 20–40% w/v MPD, 100 mM Bis-Tris pH 6.5–7.5. The crystals were flash frozen in liquid nitrogen using the mother liquor as the cryoprotectant. For the SeMet crystals, the oxidation protocol described in Sharff *et al.*[22] was used before crystallization. The oxidized SeMet crystals were further flash-soaked in 0.1% H$_2$O$_2$ in 50% w/v MPD, 100 mM Bis-Tris, pH 7.0 before flash freezing for data collection. Data collection was performed at the beamline ID14-4 at European Synchrotron Radiation Facility (Grenoble, France) for the native data set and at beamline PROXIMA 1 at the Soleil Synchrotron (Gif sur Yvette, France) for the oxidized SeMet SAD data set. Data reduction was performed using the XDS package[23]. Data collection statistics are presented in Supplementary Table 1.

Structure determination. The structure was solved by molecular replacement using the native data set and the known 20 Å resolution cryo-EM map of the trypsin-cleaved core complex as search model (Supplementary Fig. 1b)[12]. The orientation and position of the particle was determined using AMoRe[24]. The 35–20 Å resolution range was used and yielded a molecular replacement solution with a correlation coefficient of 0.39. Fourteenfold non-crystallographic symmetry averaging with RAVE, MAMA and other programs of the Uppsala suite[25–27], in combination with the CCP4 suite[28], was used for phase extension to 2.8 Å resolution using the native data set. This yielded a readily interpretable map with well-defined side chains (Supplementary Fig. 1c and Supplementary Table 1). The three component proteins TraF$_{CT}$, TraO$_{CT}$ and TraN were traced manually with the program Coot[29]. Subsequently, this structure was used to locate the selenium atoms in a SAD data set collected to 2.6 Å resolution. Refinement of these sites and phasing (PHASER[30]), followed by density modification exploiting the 14-fold non-crystallographic symmetry in PARROT[28], yielded a readily interpretable electron density map. After initial tracing by BUCCANEER[28], the partial model was used for subsequent cycles of PHASER, PARROT and BUCCANNEER. Further restrained refinement with REFMAC[31] and PHENIX[32,33] was performed using tight restraints on non-crystallographic symmetry throughout the refinement. Final stages of refinement with the map from the SAD data set showed a density for a modified cysteine in the lipoprotein TraN with small disordered lipid chains attached. However, as the density was not good enough for unambiguous building, we refrained from modelling this modification in the final model. The final model has 94.8% residues in the most favoured region, 4.7% residues in the

additionally allowed regions and 0.5% residues in the disallowed region of the Ramachandran plot. Refinement statistics are presented in the Supplementary Table 1.

Extracellular localization of a Flag-tag inserted between the α2 and α3 helices of TraF. The Flag epitope was introduced by PCR at position 332 of TraF/VirB10 in the IBA3c:*traN-traF$_{C-ST}$* plasmid[12], yielding IBA3c:*traN-traF$_{Flag,C-ST}$*. The primer sequences were: 5′-GACGACGACAAGATTCAGTACAACAGCACAGAA-3′ (forward) and 5′-GTCCTTGTAGTCGTTATTACTCTGCGTCTGGTT-3′ (reverse). Expression and purification of the resulting Flag-tagged T4S system core complex yielded a complex very similar in molecular mass to the wild-type complex (1.05 MDa), indicating that the Flag-tag does not disturb core complex assembly. *Escherichia coli* TOP10 cells containing either the IBA3c:*traN-traF$_{C-ST}$* plasmid (the expression of which results in the production of the wild-type T4S system core complex)[12] or the IBA3c:*traN-traF$_{Flag,C-ST}$* plasmid (the expression of which results in the production of the T4S system core complex containing a Flag-tag at position 332 of TraF) were grown at 37 °C to an optical density (OD$_{600 nm}$) of 0.6. Protein expression was induced by addition of 200 μg l^{-1} of anhydrotetracyclin and cells were incubated at 16 °C overnight. Cells were collected by centrifugation, resuspended in ice-cold 4% formaldehyde and incubated on ice for 15 min. The fixed cells were washed twice with PBS and re-suspended in PBS. The cell suspensions were applied to poly-L-lysine-treated microscope slides and incubated at room temperature for 20 min. The slides were rinsed in PBS and blocked in 1% goat serum-PBS for 45 min. The primary antibody (anti-Flag M2 Mab (Sigma) 1:1,000 dilution) was applied and incubated on the slide at 4 °C overnight. The slides were washed twice with PBS at room temperature and the secondary antibody (goat anti-mouse IgG1 Texas Red (Southern Biotech), 1:500 dilution) was applied and incubated on the slide for 1 h. The slides were then washed twice with PBS and were mounted using DAKO fluorescent mounting medium. The fluorescence microscopy was performed on a Zeiss Axioskop microscope, objective ×100 with oil. The images were collected using a Hamamatsu Orca ER camera.

22. Sharff, A. J., Koronakis, E., Luisi, B. & Koronakis, V. Oxidation of selenomethionine: some MADness in the method! *Acta Crystallogr. D* **56**, 785–788 (2000).
23. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795–800 (1993).
24. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
25. Jones, T. A. in *Molecular Replacement* 91–95 (CCP4 Proceedings, 1992) ⟨http://epubs.cclrc.ac.uk/bitstream/948/DL-SCI-R33.pdf⟩.
26. Kjeldgaard, M. & Jones, T. A. in *From First Map to Final Model* 59–66 (CCP4 Proceedings, 1994) ⟨http://epubs.cclrc.ac.uk/bitstream/950/DL-SCI-R35.pdf⟩.
27. Kleywegt, G. & Jones, T. Software for handling macromolecular envelopes. *Acta Crystallogr. D* **55**, 941–944 (1999).
28. Collaborative Computational Project, Number 4. The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
29. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
30. McCoy, A. *et al.* Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
31. Murshudov, G., Vagin, A. & Dodson, E. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
32. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
33. Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. The Phenix refinement framework. *CCP4 Newsl.* **42**, contribution 8 (2005) ⟨http://www.ccp4.ac.uk/newsletters/newsletter42.pdf⟩.

ARTICLES

The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes

Lisa R. Racki[1]*, Janet G. Yang[1]*, Nariman Naber[1], Peretz D. Partensky[1], Ashley Acevedo[1], Thomas J. Purcell[1], Roger Cooke[1], Yifan Cheng[1,2] & Geeta J. Narlikar[1]

Evenly spaced nucleosomes directly correlate with condensed chromatin and gene silencing. The ATP-dependent chromatin assembly factor (ACF) forms such structures *in vitro* and is required for silencing *in vivo*. ACF generates and maintains nucleosome spacing by constantly moving a nucleosome towards the longer flanking DNA faster than the shorter flanking DNA. How the enzyme rapidly moves back and forth between both sides of a nucleosome to accomplish bidirectional movement is unknown. Here we show that nucleosome movement depends cooperatively on two ACF molecules, indicating that ACF functions as a dimer of ATPases. Further, the nucleotide state determines whether the dimer closely engages one or both sides of the nucleosome. Three-dimensional reconstruction by single-particle electron microscopy of the ATPase–nucleosome complex in an activated ATP state reveals a dimer architecture in which the two ATPases face each other. Our results indicate a model in which the two ATPases work in a coordinated manner, taking turns to engage either side of a nucleosome, thereby allowing processive bidirectional movement. This novel dimeric motor mechanism differs from that of dimeric motors such as kinesin and dimeric helicases that processively translocate unidirectionally and reflects the unique challenges faced by motors that move nucleosomes.

Chromatin remodelling motors have essential roles in organizing the chromatin state for regulating eukaryotic genomes, yet how they carry out their myriad activities is poorly understood. Their substrate, the nucleosome, contains 147 base pairs (bp) of DNA wrapped in ~1.5 turns around an octamer of histone proteins. Even the smallest movement of the histone octamer relative to the DNA presumably requires a coordinated process of breaking and reforming the many histone–DNA contacts. The ACF chromatin remodelling complex exemplifies the task, as it is able to move nucleosomes to create evenly spaced nucleosomal arrays that contain equal DNA on either side of each nucleosome[1–10]. These evenly spaced arrays are important for packaging the underlying DNA into silent chromatin structures *in vivo*[1–10].

ACF is part of the ISWI family of remodelling complexes. The ATPase subunits of ISWI complexes can move nucleosomes by themselves while the accessory subunits modulate this basic activity[11–15]. The human ACF complex consists of one ATPase subunit, SNF2h, and one accessory subunit, Acf1 (also known as SMARCA5 and BAZ1A, respectively)[6,7]. SNF2h is part of the SF2 family of DExx box proteins that includes helicases and nucleic acid translocases[16]. The ATPase domain of SNF2h has two RecA-like domains, which are thought to form a cleft within which ATP binds. SNF2h also has an α-helical extension comprised of three additional domains, HAND, SANT and SLIDE, which are thought to have a role in binding flanking DNA[17,18]. We showed previously that ACF generates a dynamic equilibrium in which nucleosomes with equal flanking DNA on either side accumulate[8]. Our data showed that ACF achieves the dynamic equilibrium by constantly sampling either side of the nucleosome. This sampling mechanism raised the question of how ACF efficiently switches back and forth between both sides of a nucleosome. We hypothesized that understanding how the ATP state affects interactions of the enzyme with the nucleosome would provide insight into the sampling process.

Previous work has shown that ISWI enzymes require a basic patch, $K_{16}R_{17}H_{18}R_{19}$, on the amino-terminal tail of histone H4 for maximal activity[19–25]. The role of the H4 tail is not known, but it has been hypothesized that an acidic patch on the ATPase domain of ISWI enzymes may interact with the basic patch on the H4 tail[17]. These previous observations suggest that the ATPase subunit contacts the H4 tail and that the contacts may change during the ATPase cycle. We therefore used changes in the mobility of the H4 tail as a handle to follow how changes in the nucleotide state alter interactions between SNF2h and the nucleosome. We used electron paramagnetic resonance (EPR) spectroscopy for these studies[26]. We covalently attached a maleimide spin probe to a cysteine introduced in place of an alanine at position 15 on the H4 tail (A15C-MSL, Supplementary Fig. 1b), which is directly adjacent to the basic patch. Thermal fluctuations cause a spin label attached to a protein to undergo motion in a spatial region defined by the adjacent protein surface. The resulting EPR spectrum is a highly sensitive measure of the region accessible to the probe. Conformational changes can thus be detected via changes in probe mobility, and these are monitored as changes in the EPR spectrum. EPR can also resolve and quantify multiple states and is particularly powerful in monitoring transitions between unstructured and structured regions of proteins. The A15C-MSL nucleosomes were assembled using an asymmetric DNA template comprising the 601 positioning sequence with 60 bp of flanking DNA on one side (0–601–60, Fig. 1a and Supplementary Fig. 1a)[27]. The presence of the probe did not alter the maximal rate of nucleosome remodelling by SNF2h (data not shown).

In the absence of SNF2h, the EPR spectrum of the A15C-MSL probe indicated a highly mobile probe (Fig. 1a, top spectrum). The high mobility of the probe in unbound nucleosomes suggested that the H4 N-terminal tails are largely unstructured. Next, we determined how binding of SNF2h altered the mobility of the H4 tail.

[1]Department of Biochemistry and Biophysics, University of California, 600 16th Street, [2]The W.M. Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California 94158, USA.
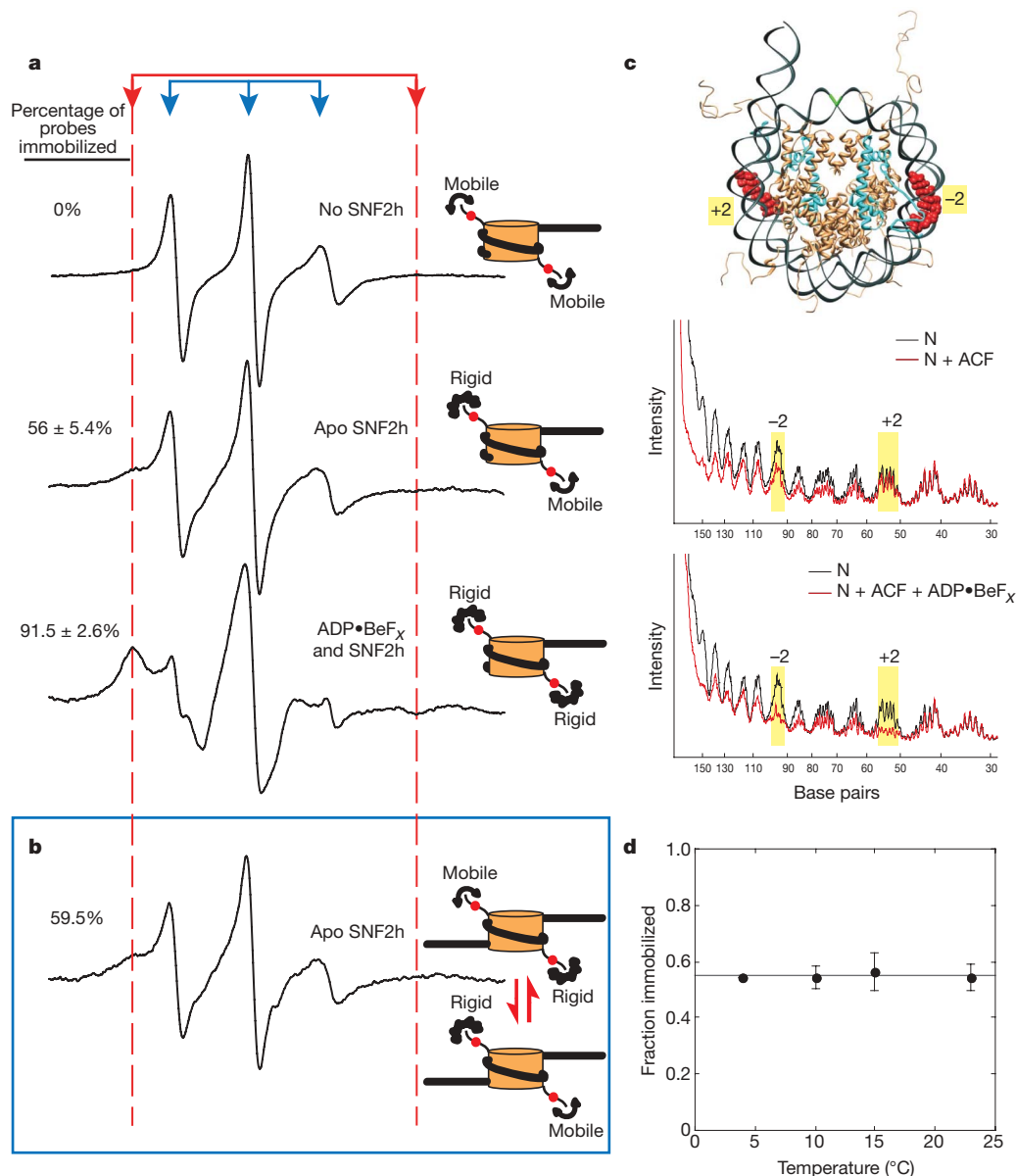*These authors contributed equally to this work

Figure 1 | ATP state regulates immobilization of the histone H4 tail and proximal interactions. **a**, Left panels: EPR spectra of MSL labelled 0–601–60 nucleosomes; right panels: schematic interpretation of EPR spectra, on the basis of data from **a**, **c** and **d**. Binding of apo SNF2h to the nucleosome decreases the mobility of half the H4 tails. SNF2h binding in the presence of ADP•BeF$_x$ decreases the mobility of both H4 tails. **b**, EPR spectrum of apo SNF2h bound to spin-labelled 60–601–60 nucleosomes shows that only one of the two H4 tails is immobilized. **c**, Hydroxyl radical foot-printing of ACF on 0–601–60 nucleosomes. Top panel: schematic of mononucleosome structure with 12 bp of flanking DNA on one side, with dyad in green, histone H4 in blue, and the region surrounding SHL(−2) and (+2) in red. Middle panel: protection patterns for nucleosomes alone (black line, N) compared to nucleosomes bound by apo-ACF (red line, N + ACF). Bottom panel: nucleosomes alone (black line, N) compared to nucleosomes bound by ACF in the presence of ADP•BeF$_x$ (red line, N + ACF + ADP•BeF$_x$). Yellow bars highlight protection in the SHL(−2) and (+2) regions. **d**, Temperature dependence of probe immobilization in the apo-SNF2h–nucleosome complex. Slope of the straight line = 0 within experimental error. Error bars represents s.e.m.

When the nucleosomes were saturated with SNF2h in the absence of nucleotide (apo state), the EPR spectrum showed two sets of spectral components as indicated by the arrows (Fig. 1a, middle spectrum). The inner spectral components (blue arrows) are indicative of a highly mobile probe whereas the wider set of spectral components (high-lighted by the red dashed lines) and the broadening of the central peak are indicative of a second state with more restricted mobility. A given peak height in the left-most immobilized spectral component represents 4.1 times more spins than the same peak height for the mobile component. Deconvolution of the spectra indicated that about half of the H4 tails were in each of the two states (56 ± 5.4% in the immobilized state, see Supplementary Fig. 2 for fitting and quantification method)[28]. In the presence of ADP, the immobilized subpopulation also constituted half of the probes (spectra not shown). Our attempts

to trap the SNF2h–nucleosome complex in the ATP state using ATP analogues were unsuccessful as these analogues either supported low levels of remodelling (β-γ-imidoadenosine 5′-phosphate (AMP-PNP), ATPγS) or did not detectably inhibit remodelling (adenosine 5′-monophosphate with (phosphonomethyl)phosphonic acid, AMP-PCP). We were however able to mimic an activated ATP state using the analogue ADP•BeF$_x$. In contrast to the data in the apo state and with ADP, almost all of the probe on the H4 tail became immobilized in the presence of SNF2h and ADP•BeF$_x$ (91.5 ± 2.6% of probe in the immobilized peak). This change is shown by the increase in spectral intensity of the immobilized component (left-most peak, Fig. 1a, bottom spectrum). This marked increase in the amount of probes immobilized suggested that both H4 tails were immobilized by SNF2h in the presence of ADP•BeF$_x$. Together, these data indicate

that SNF2h induces nucleotide-dependent changes in the H4 tail conformation such that in the apo and ADP states, half the H4 tails are immobilized, and in an activated ATP state mimicked by ADP•BeF$_x$, all H4 tails are immobilized.

The EPR data raised two possibilities for how SNF2h binds the nucleosome in the apo and ADP states: (1) SNF2h symmetrically binds both H4 tails and each H4 tail exists in a two-state equilibrium between mobile and immobile states (with an equilibrium constant of 1), or (2) SNF2h asymmetrically binds only one of the two H4 tails. For model (1), we expect immobilization to increase when lowering temperature as the highly mobile state is entropically favoured, whereas the structured immobile state is enthalpically favoured, as seen for docking of the kinesin neck linker[29]. The fraction of H4 tails immobilized was unchanged, within error, from 23 °C to 4 °C (Fig. 1d, 54.6% immobilized at 4 °C and 23 °C). A van 't Hoff plot of the equilibrium constant for H4 tail mobility as a function of temperature yields a ΔH of 0.76 kJ mol^{-1}, and ΔS of 4.4×10^{-3} kJ mol^{-1} K^{-1} (Supplementary Fig. 3), values that are substantially smaller than the favourable ΔH of 50 kJ mol^{-1} and unfavourable ΔS of 0.17 kJ mol^{-1} K^{-1} for docking of the kinesin neck linker[29]. These data rule out model (1) and provide strong support for the asymmetric binding of model (2). To test model (2) further, we used hydroxyl radical footprinting of the same nucleosome construct to follow changes in ACF contacts as a function of nucleotide state (Fig. 1c). In the apo state, ACF binding induces asymmetric protection of nucleosomal DNA: protection is observed in the SHL(−2) region, but not the SHL(+2) region, consistent with other ISWI complexes and with model (2) (refs 9, 30). In contrast, in the ADP•BeF$_x$ state, ACF binding results in significant protection in both SHL(−2) and SHL(+2) regions, consistent with the EPR data.

The asymmetry with respect to H4 tail binding observed in the apo state could arise either because of the presence of asymmetric flanking DNA, or because of structural constraints placed by the apo state. To distinguish between these possibilities we repeated the EPR experiment using nucleosomes with 60 bp of flanking DNA on both sides (60–601–60 template). Apo-SNF2h still immobilized only one of the two H4 tails in the context of this symmetric nucleosome (Fig. 1b). These data strongly support a model in which apo-SNF2h can only bind one H4 tail at a time, and the availability of flanking DNA biases which side of the nucleosome the enzyme binds preferentially. Together the above data indicate that the enzyme switches between an asymmetric conformation where it interacts with one H4 tail at a time in the apo state and a more symmetric conformation where it binds both H4 tails in the ADP•BeF$_x$ state.

The observation that SNF2h binds both H4 tails in the presence of ADP•BeF$_x$ suggests that either one SNF2h molecule bridges both H4 tails, or SNF2h binds as a dimer such that each ATPase contacts an H4 tail. To distinguish between these models we first investigated the oligomeric state of SNF2h alone. Using equilibrium analytical ultracentrifugation we found that unbound SNF2h is a monomer (data not shown). Because several well-studied dimeric helicases dimerize on binding their DNA substrates, we next determined whether SNF2h dimerizes on nucleosomes[31]. If dimerization of SNF2h is tightly coupled to nucleosome binding we expected to see cooperative SNF2h binding. We measured the binding to nucleosomes by taking advantage of our observation that the fluorescence of a Cy3 dye attached near the entry site of the DNA increases upon SNF2h binding (Fig. 2a). We find that in the apo state, SNF2h binds to the nucleosome cooperatively, consistent with previous observations of cooperative binding by the *Drosophila* ISWI protein[32]. The Hill coefficient of 1.8 indicates that at least two molecules of SNF2h bind in a manner such that binding of one molecule is strongly coupled to binding of the second (Fig. 2b). The EPR data from Fig. 1 and the Hill coefficient of 1.8 together indicate that in the apo state, SNF2h binds as a dimer, but only one of the two SNF2h molecules engages an H4 tail.

To determine whether two SNF2h molecules were necessary to mediate maximal nucleosome remodelling, we measured the dependence

of chromatin remodelling activity on SNF2h concentration using a fluorescence resonance energy transfer (FRET)-based method (Fig. 2c). The rate constant of remodelling also depends cooperatively on SNF2h concentration with a Hill coefficient of 1.8 (Fig. 2d, left panel). We next
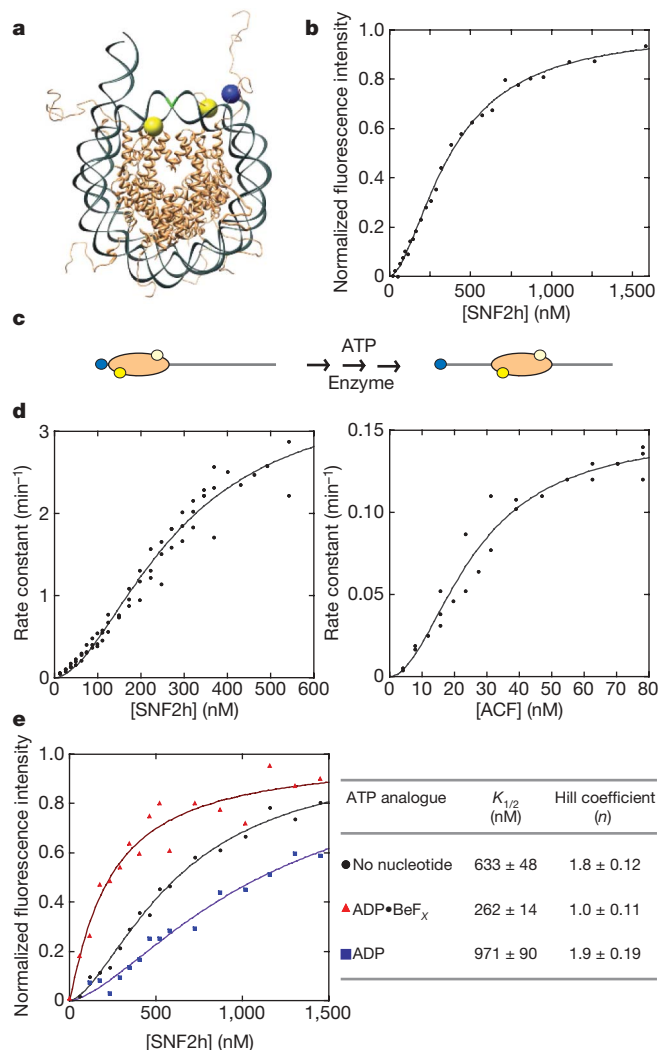
Figure 2 | SNF2h and ACF function as dimers of ATPases. a, Schematic of nucleosome structure with dye attachment sites for **b** and **d**. The DNA is end-labelled with Cy3 (blue) on the shorter flanking DNA. The octamer is labelled with Cy5 at H2A-120C (yellow). **b**, Cy3 fluorescence intensity of the construct shown in **a** as a function of SNF2h concentration, using nucleosomes with 78 bp of flanking DNA on one side. A representative replicate curve is shown. Data are fit to the general equation for cooperative binding (see Methods). Hill coefficient $n = 1.8 \pm 0.17$; $K_{1/2} = 353 \pm 30$ nM. **c**, Schematic of FRET-based nucleosome remodelling assay. Rate constant of remodelling is measured by following the decrease in FRET between Cy3 and Cy5 in the presence of ATP. **d**, Left panel: nucleosome remodelling rate constant as a function of SNF2h concentration for nucleosomes with 78 bp of flanking DNA; right panel, nucleosome remodelling rate constant as a function of ACF concentration for nucleosomes with 20 bp of flanking DNA. Hill coefficient $n = 1.8 \pm 0.1$; $K'_{1/2} = 281 \pm 32$ nM for SNF2h and Hill coefficient $n = 1.9 \pm 0.3$; $K'_{1/2} = 26 \pm 3$ nM for ACF. Each panel represents global fits to data obtained from three independent experiments. **e**, SNF2h binds as a cooperative dimer to the nucleosome in the absence of nucleotide (black circles), and in the presence of ADP (blue squares). In the presence of ADP•BeF$_x$, SNF2h binds non-cooperatively (red triangles). These binding measurements were carried out with nucleosomes containing 40 bp of flanking DNA on one side and a Cy3 label on the short DNA end. Binding of SNF2h to these nucleosomes is ~twofold weaker relative to the nucleosomes used in **b**[9,47]. A representative replicate curve with each nucleotide analogue is shown (left panel), and the average $K_{1/2}$ and Hill coefficient from three replicates is shown (table). Errors represent s.e.m.

ATP analogue	$K_{1/2}$ (nM)	Hill coefficient (n)
● No nucleotide	633 ± 48	1.8 ± 0.12
▲ ADP•BeF$_X$	262 ± 14	1.0 ± 0.11
■ ADP	971 ± 90	1.9 ± 0.19

determined whether the entire ACF complex also functions most effectively as a dimer. We analogously saw a cooperative dependence of the remodelling rate constant on ACF concentration with a Hill coefficient of 1.9 (Fig. 2d, right panel). Together, these data strongly indicate that the predominant functional form of ACF is a dimer of ATPases.

A hallmark of dimeric motors such as kinesin and the *Escherichia coli* Rep helicase is that they cycle between states in which one motor subunit is engaged with the substrate and states in which both motor subunits are transiently engaged[31]. By working in coordinated pairs, one motor subunit can serve as an anchor to the other moving motor to prevent dissociation from the substrate. The ATP state helps regulate the affinity of the motor for the substrate. Our EPR results show that a SNF2h dimer analogously cycles between at least two states, one in which only one ATPase engages an H4 tail and another in which both ATPases engage the two H4 tails. To determine if these different conformational states reflect states with different affinities, we measured the affinity of SNF2h for the nucleosome in different ATP states (Fig. 2e). In the presence of ADP, SNF2h bound cooperatively but with slightly weaker affinity than in the apo state. In both the apo and ADP states, the high cooperativity of binding indicates that binding of one SNF2h molecule by itself is very weak and requires the presence of another SNF2h molecule to increase its overall affinity. The cooperativity could arise either from direct SNF2h–SNF2h contacts or could be mediated through a conformational change in the nucleosome without direct SNF2h–SNF2h contacts. In the presence of ADP•BeF$_x$, the $K_{1/2}$ for SNF2h binding is approximately threefold lower than that in the apo state, indicating a stronger binding affinity. Further, SNF2h binding in the presence of ADP•BeF$_x$ is not cooperative (Hill coefficient = 1). These data show that in presence of ADP•BeF$_x$ the affinity of each SNF2h molecule is sufficiently high such that binding of one SNF2h molecule is no longer highly dependent on the presence of the other.

Our finding that two ATPases are required for maximal nucleosome remodelling raises the question of how the nucleosome structure accommodates two SNF2h molecules. Other dimeric motors, such as kinesin and helicases, are oriented such that each ATPase subunit can take turns translocating on the polymeric substrate in the same direction[31]. We used negative stain electron microscopy to visualize the complex of SNF2h with nucleosomes in the presence of ADP•BeF$_x$ (Fig. 3, Supplementary Fig. 4 and Supplementary Methods). Nucleosomes with 60 bp of flanking DNA (0–601–60) were incubated with SNF2h concentrations comparable to the $K_{1/2}$ for nucleosomes in the presence of ADP•BeF$_x$, adsorbed to a glow-discharged carbon film, and negatively stained with uranyl formate. The specimen was imaged at tilt angles of 60° and 0° (Supplementary Fig. 5). A total of 10,059 pairs of particles were interactively selected from 100 image pairs. Classification of particles from images of untilted specimen shows three distinct classes (Fig. 3a, b) that can be clearly recognized as a nucleosome by itself and a nucleosome with either one or two SNF2h molecules bound. On average ~70% of the complexes contained two SNF2h molecules bound. The singly bound SNF2h molecules could reflect the use of non-saturating SNF2h, a technical necessity to prevent particle crowding on the grid. In the complexes with two SNF2h molecules bound, the flanking DNA was not clearly visible in the two-dimensional (2D) class averages, possibly because the flanking DNA is flexible and gets averaged out. An alternative possibility is that in most of the complexes, the flanking DNA is rearranged due to interaction with a domain of SNF2h. Consistent with this possibility we do not clearly observe the extended HAND-SANT-SLIDE domain that has been shown to interact with flanking DNA in the apo state[17]. We hypothesize that there may be a conformational rearrangement of the HAND-SANT-SLIDE domain in the presence of ADP•BeF$_x$. Further, no large region of direct contact between the two SNF2h molecules is apparent, consistent with the Hill coefficient of 1.0 in this state (Fig. 2e).

In the 2D class averages, the region of each SNF2h monomer that interacts with the nucleosome appears to contain two globular lobes

Figure 3 | Visualization of SNF2h bound to the nucleosome in the presence of ADP•BeF$_x$ using electron microscopy. a, Three different views of the 3D reconstruction of dimeric SNF2h bound to the nucleosome (left panels) and corresponding representative 2D class averages (right panels). The crystal structure of the core mononucleosome was placed manually into the 3D reconstruction. Histone H4 is highlighted in red. The isosurface of the 3D reconstruction at high threshold is shown in blue, and low threshold in grey. **b**, Left panel, representative 2D class average of negative stain electron microscopy images of unbound nucleosomes; right panel, representative 2D class average of one SNF2h bound to a nucleosome. Numbers used to calculate a particular class average shown in lower left corner in **a** and **b**. **c**, Representative 2D class averages of SNF2h alone.

(Fig. 3a). These lobes may represent the two RecA-related ATP binding folds observed in SF2 family motors[33,34]. The two lobes are also apparent in the 2D class averages of SNF2h alone (Fig. 3c and Supplementary Fig. 6). Three-dimensional (3D) reconstructions of the nucleosome with two (or one) SNF2h bound were calculated using the well-established random conical tilt approach to a resolution of ~27 Å without the explicit application of any two-fold symmetry (Fig. 3a and Supplementary Figs 5 and 7)[35]. The two SNF2h molecules face each other on the nucleosome, and seem to obey the two-fold symmetry of the nucleosome with one putative ATPase domain at SHL(+2) and one at SHL(−2) (Fig. 3a, two-fold or opposing symmetry most apparent in the middle panel). Consistent with the EPR and footprinting data in Fig. 1, each SNF2h monomer seems to directly contact one H4 N-terminal tail and the SHL(−2)/(+2) regions in the ADP•BeF$_x$ state. Although the overall architecture appears almost symmetric, given the 27 Å resolution, any local structural asymmetries that may exist between the two SNF2h molecules cannot be resolved.

The above architecture raises the question of how the dimeric partners cooperate rather than compete in a 'tug of war'. Our findings indicate an 'alternating action' model schematized in Fig. 4. In this model, each ATPase takes turns in engaging the flanking DNA on either side and the corresponding H4 tail at SHL(−2) or (+2) (refs 9, 17, 30). An ability of the two ATPases to take turns, as indicated by our observation that the apo state of the enzyme engages only one H4 tail at a time, would help avoid a tug of war situation. The ATPase that engages the longer DNA hydrolyses ATP faster, as shown previously[8,36–38]. This ATPase becomes the leading ATPase and sets the direction of nucleosome movement by translocating on nucleosomal DNA[9,37,38]. The leading ATPase generates a DNA loop/wave that can propagate across the histone octamer as suggested previously[39–41]. The second, subordinate ATPase could then further act as another anchor to stabilize the intermediate while the leading ATPase is translocating (Fig. 4, mimicked by ADP•BeF$_x$). Interaction with the second H4 tail may help in the binding of the subordinate ATPase. In the simplest version of this model, the subordinate ATPase does not bind or hydrolyse ATP once the leading ATPase fires[42]. This division of labour between identical subunits is analogous to hexameric helicases where occupancy of one ATPase subunit regulates the affinity of an adjacent subunit for nucleotide[43]. Whether the communication between the two ATPases is direct or through the nucleosome remains an important future question. A variation of this model in which the non-leading ATPase also hydrolyses ATP is described in Supplementary Fig. 8. Successive rounds of sampling and translocation would then equalize the DNA on either side of a nucleosome. A dimer-based mechanism is also indicated by single-molecule data

showing that dimeric ACF complexes can switch the direction of nucleosome translocation several times without dissociation[44]. Our results help explain the significance of previous observations that two *Drosophila* ACF molecules can bind in the context of DNA and provide a mechanistic explanation for the processive action of ISWI complexes[40,45,46].

We hypothesize that in contrast to kinesin and dimeric helicases, whose biological functions require unidirectional translocation along a largely uniform polymeric substrate, the biological functions of chromatin remodelling enzymes like ACF place very different demands on motor architecture. The opposing architecture of the two motors in ACF may enable ACF to rapidly and processively change the direction of nucleosome movement in order to achieve a defined spacing. It will be interesting to investigate whether bidirectional movement via dimerization is a general feature of enzymes that space nucleosomes, and whether remodelling enzymes with other activities use different strategies.

METHODS SUMMARY

EPR measurements were performed with an EMX EPR spectrometer from Bruker Instruments. First derivative, X-band spectra were recorded in a high-sensitivity microwave cavity using 50 s scans, 100 Gauss wide magnetic field sweeps. Electron microscopy samples were adsorbed to a glow-discharged copper grid coated with carbon film for 30 s followed by conventional negative stain with 0.75% uranyl formate. Images were collected using a Tecnai T12 microscope (FEI) and recorded at a magnification of ×52,000 with an UltraScan 4096 × 4096 pixel CCD camera (Gatan). Full methods are described in Supplementary Methods.

1. Corona, D. F. *et al.* ISWI regulates higher-order chromatin structure and histone H1 assembly *in vivo*. *PLoS Biol.* **5**, e232 (2007).
2. Fyodorov, D. V., Blower, M. D., Karpen, G. H. & Kadonaga, J. T. Acf1 confers unique activities to ACF/CHRAC and promotes the formation rather than disruption of chromatin *in vivo*. *Genes Dev.* **18**, 170–183 (2004).
3. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
4. Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388**, 598–602 (1997).
5. Deuring, R. *et al.* The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure *in vivo*. *Mol. Cell* **5**, 355–365 (2000).
6. Poot, R. A. *et al.* HuCHRAC, a human ISWI chromatin remodelling complex contains hACF1 and two novel histone-fold proteins. *EMBO J.* **19**, 3377–3387 (2000).
7. Bochar, D. A. *et al.* A family of chromatin remodeling factors related to Williams syndrome transcription factor. *Proc. Natl Acad. Sci. USA* **97**, 1038–1043 (2000).
8. Yang, J. G., Madrid, T. S., Sevastopoulos, E. & Narlikar, G. J. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nature Struct. Mol. Biol.* **13**, 1078–1083 (2006).
9. Kagalwala, M. N., Glaus, B. J., Dang, W., Zofall, M. & Bartholomew, B. Topography of the ISW2-nucleosome complex: insights into nucleosome spacing and chromatin remodeling. *EMBO J.* **23**, 2092–2104 (2004).
10. Sun, F. L., Cuaycong, M. H. & Elgin, S. C. Long-range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. *Mol. Cell. Biol.* **21**, 2867–2879 (2001).
11. Corona, D. F. *et al.* ISWI is an ATP-dependent nucleosome remodeling factor. *Mol. Cell* **3**, 239–245 (1999).
12. Aalfs, J. D., Narlikar, G. J. & Kingston, R. E. Functional differences between the human ATP-dependent nucleosome remodeling proteins BRG1 and SNF2H. *J. Biol. Chem.* **276**, 34270–34278 (2001).
13. Eberharter, A. *et al.* Acf1, the largest subunit of CHRAC, regulates ISWI-induced nucleosome remodeling. *EMBO J.* **20**, 3781–3788 (2001).
14. Längst, G., Bonte, E. J., Corona, D. F. & Becker, P. B. Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. *Cell* **97**, 843–852 (1999).
15. Ito, T. *et al.* ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. *Genes Dev.* **13**, 1529–1539 (1999).
16. Dürr, H., Flaus, A., Owen-Hughes, T. & Hopfner, K. P. Snf2 family ATPases and DExx box helicases: differences and unifying concepts from high-resolution crystal structures. *Nucleic Acids Res.* **34**, 4160–4167 (2006).
17. Dang, W. & Bartholomew, B. Domain architecture of the catalytic subunit in the ISW2-nucleosome complex. *Mol. Cell. Biol.* **27**, 8306–8317 (2007).

Figure 4 | Simple model for how a dimeric ACF moves nucleosomes. H4 tail is in red and the two ATPases in ACF are shown in blue and purple. Only one subunit binds ATP at a time. In the ATP state, each ATPase subunit takes turns in binding the flanking DNA. The ATPase that binds the longer flanking DNA (purple) hydrolyses ATP faster and starts translocating DNA across the nucleosome. During and after hydrolysis, the second ATPase (blue) also engages the nucleosome, preventing loss of the DNA loop-containing intermediate (mimicked by ADP•BeF$_x$). In the ADP state, the non-translocating monomer disengages and the translocating monomer remains engaged with the nucleosome. ATP state may also regulate the extent of any direct contacts between the two ATPase subunits and such contacts may be substantially fewer in the ADP•BeF$_x$ state than in other ATP states.

18. Grüne, T. *et al.* Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI. *Mol. Cell* **12**, 449–460 (2003).
19. Clapier, C. R., Langst, G., Corona, D. F., Becker, P. B. & Nightingale, K. P. Critical role for the histone H4 N terminus in nucleosome remodeling by ISWI. *Mol. Cell. Biol.* **21**, 875–883 (2001).
20. Clapier, C. R., Nightingale, K. P. & Becker, P. B. A critical epitope for substrate recognition by the nucleosome remodeling ATPase ISWI. *Nucleic Acids Res.* **30**, 649–655 (2002).
21. Hamiche, A., Kang, J. G., Dennis, C., Xiao, H. & Wu, C. Histone tails modulate nucleosome mobility and regulate ATP-dependent nucleosome sliding by NURF. *Proc. Natl Acad. Sci. USA* **98**, 14316–14321 (2001).
22. Fazzio, T. G., Gelbart, M. E. & Tsukiyama, T. Two distinct mechanisms of chromatin interaction by the Isw2 chromatin remodeling complex *in vivo*. *Mol. Cell. Biol.* **25**, 9165–9174 (2005).
23. Ferreira, H., Flaus, A. & Owen-Hughes, T. Histone modifications influence the action of Snf2 family remodelling enzymes by different mechanisms. *J. Mol. Biol.* **374**, 563–579 (2007).
24. Dang, W., Kagalwala, M. N. & Bartholomew, B. Regulation of ISW2 by concerted action of the histone H4 tail and extranucleosomal DNA. *Mol. Cell. Biol.* **26**, 7388–7396 (2006).
25. Shogren-Knaak, M. *et al.* Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**, 844–847 (2006).
26. Rice, S. *et al.* A structural change in the kinesin motor protein that drives motility. *Nature* **402**, 778–784 (1999).
27. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
28. Naber, N., Purcell, T. J., Pate, E. & Cooke, R. Dynamics of the nucleotide pocket of myosin measured by spin-labeled nucleotides. *Biophys. J.* **92**, 172–184 (2007).
29. Rice, S. *et al.* Thermodynamic properties of the kinesin neck-region docking to the catalytic core. *Biophys. J.* **84**, 1844–1854 (2003).
30. Schwanbeck, R., Xiao, H. & Wu, C. Spatial contacts and nucleosome step movements induced by the NURF chromatin remodeling complex. *J. Biol. Chem.* **279**, 39933–39941 (2004).
31. Lohman, T. M., Thorn, K. & Vale, R. D. Staying on track: common features of DNA helicases and microtubule motors. *Cell* **93**, 9–12 (1998).
32. Chin, J., Langst, G., Becker, P. B. & Widom, J. Fluorescence anisotropy assays for analysis of ISWI-DNA and ISWI-nucleosome interactions. *Methods Enzymol.* **376**, 3–16 (2003).
33. Dürr, H., Korner, C., Muller, M., Hickmann, V. & Hopfner, K. P. X-ray structures of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core and its complex with DNA. *Cell* **121**, 363–373 (2005).
34. Thomä, N. H. *et al.* Structure of the SWI2/SNF2 chromatin-remodeling domain of eukaryotic Rad54. *Nature Struct. Mol. Biol.* **12**, 350–356 (2005).
35. Radermacher, M. *et al.* Cryo-electron microscopy and three-dimensional reconstruction of the calcium release channel/ryanodine receptor from skeletal muscle. *J. Cell Biol.* **127**, 411–423 (1994).
36. Stockdale, C., Flaus, A., Ferreira, H. & Owen-Hughes, T. Analysis of nucleosome repositioning by yeast ISWI and Chd1 chromatin remodeling complexes. *J. Biol. Chem.* **281**, 16279–16288 (2006).
37. Zofall, M., Persinger, J., Kassabov, S. R. & Bartholomew, B. Chromatin remodeling by ISW2 and SWI/SNF requires DNA translocation inside the nucleosome. *Nature Struct. Mol. Biol.* **13**, 339–346 (2006).
38. Whitehouse, I., Stockdale, C., Flaus, A., Szczelkun, M. D. & Owen-Hughes, T. Evidence for DNA translocation by the ISWI chromatin-remodeling enzyme. *Mol. Cell. Biol.* **23**, 1935–1945 (2003).
39. Cairns, B. R. Chromatin remodeling: insights and intrigue from single-molecule studies. *Nature Struct. Mol. Biol.* **14**, 989–996 (2007).
40. Strohner, R. *et al.* A 'loop recapture' mechanism for ACF-dependent nucleosome remodeling. *Nature Struct. Mol. Biol.* **12**, 683–690 (2005).
41. Längst, G. & Becker, P. B. Nucleosome remodeling: one mechanism, many phenomena? *Biochim. Biophys. Acta* **1677**, 58–63 (2004).
42. Racki, L. R. & Narlikar, G. J. ATP-dependent chromatin remodeling enzymes: two heads are not better, just different. *Curr. Opin. Genet. Dev.* **18**, 137–144 (2008).
43. Enemark, E. J. & Joshua-Tor, L. On helicases and other motor proteins. *Curr. Opin. Struct. Biol.* **18**, 243–257 (2008).
44. Blosser, T. R., Yang, J. G., Stone, M. D., Narlikar, G. J. & Zhuang, X. Dynamics of nucleosome remodeling by individual ACF complexes. *Nature* doi:10.1038/nature08627 (in the press).
45. Fyodorov, D. V. & Kadonaga, J. T. Dynamics of ATP-dependent chromatin assembly by ACF. *Nature* **418**, 896–900 (2002).
46. Gangaraju, V. K., Prasad, P., Srour, A., Kagalwala, M. N. & Bartholomew, B. Conformational changes associated with template commitment in ATP-dependent chromatin remodeling by ISW2. *Mol. Cell* **35**, 58–69 (2009).
47. He, X., Fan, H. Y., Narlikar, G. J. & Kingston, R. E. Human ACF1 alters the remodeling strategy of SNF2h. *J. Biol. Chem.* **281**, 28636–28647 (2006).

ARTICLES

Dynamics of nucleosome remodelling by individual ACF complexes

Timothy R. Blosser[1,2], Janet G. Yang[3], Michael D. Stone[1,4], Geeta J. Narlikar[3] & Xiaowei Zhuang[1,4,5]

The ATP-dependent chromatin assembly and remodelling factor (ACF) functions to generate regularly spaced nucleosomes, which are required for heritable gene silencing. The mechanism by which ACF mobilizes nucleosomes remains poorly understood. Here we report a single-molecule FRET study that monitors the remodelling of individual nucleosomes by ACF in real time, revealing previously unknown remodelling intermediates and dynamics. In the presence of ACF and ATP, the nucleosomes exhibit gradual translocation along DNA interrupted by well-defined kinetic pauses that occurred after approximately seven or three to four base pairs of translocation. The binding of ACF, translocation of DNA and exiting of translocation pauses are all ATP-dependent, revealing three distinct functional roles of ATP during remodelling. At equilibrium, a continuously bound ACF complex can move the nucleosome back-and-forth many times before dissociation, indicating that ACF is a highly processive and bidirectional nucleosome translocase.

The packaging of DNA into chromatin represses essential nucleic acid transactions, such as transcription, replication, repair and recombination. This repression is in part regulated by chromatin remodelling enzymes, which couple the energy of ATP hydrolysis to the assembly and mobilization of nucleosomes. ATP-dependent chromatin remodelling enzymes can be classified into several sub-families, SWI/SNF, ISWI, CHD/Mi2 and INO80, depending on their composition and function[1–5]. Despite possessing a conserved super-family 2 ATPase subunit that facilitates DNA translocation[6,7], different subfamilies exhibit divergent remodelling activities. For example, the ISWI enzymes have been shown to translocate the histone octa-mer along DNA and generate a repositioned nucleosome with a canonical structure[8–11], whereas the SWI/SNF enzymes generate a variety of products including repositioned nucleosomes, alternative nucleosome structures containing DNA loops, and nucleosomes with altered histone composition[1–5]. The kinetic intermediates and pathways through which the nucleosome structure evolves during remodelling, however, remain largely elusive. Single molecule experi-ments are ideally suited to probe these dynamics. Recently, optical and magnetic tweezers have been used to study individual SWI/SNF remodellers, providing direct measurements of DNA translocation and loop formation by these enzymes[12–14]. In this work, we established a single-molecule fluorescence resonance energy transfer (FRET)[15–17] assay to characterize the structural dynamics and kinetic intermedi-ates of nucleosomes during remodelling. Human ACF[18–22], a repre-sentative member of the ISWI family of remodellers, was investigated using this approach.

Probing nucleosome translocation by FRET

For FRET characterizations, we labelled histone octamers with a donor dye (Cy3) on histone H2A (ref. 23) and reconstituted mono-nucleosomes with the Cy3-labelled octamer and a double-stranded DNA that contained an acceptor dye (Cy5) and a biotin at opposite ends. Unless otherwise indicated, we used the 601 nucleosome posi-tioning sequence[24] to place the octamer three base pairs (bp) away from the Cy5-labelled exit end of the DNA, leaving 78 bp of linker

DNA on the entry side (Fig. 1a, Supplementary Fig. 1a, $n = 3$ bp). The nucleosomes were then anchored to a microscope slide via a biotin–streptavidin linkage and imaged by a total-internal-reflection-fluorescence (TIRF) microscope[25]. The presence of two H2A subunits in each octamer led to a heterogeneous population of nucleosomes with three different labelling configurations: (1) donor on the H2A subunit proximal to the acceptor, (2) donor on the H2A subunit distal to the acceptor, (3) donor on both H2A subunits. Single-molecule detection allowed these configurations to be discriminated. Three distinct peaks centred at FRET = 0.88, 0.75 and 0.58 were observed in the FRET distribution (Fig. 1b). The assignment of these peaks to the three labelling configurations was further confirmed by individual FRET time traces, which showed one- or two-step photobleaching for nucleosomes bearing one or two donor dyes, respectively (Sup-plementary Fig. 2). In the following, we focus our analyses on nucleo-somes containing a single donor on the proximal H2A (FRET = 0.88) to maximize the dynamic range in our experiments.

Recombinant ACF, comprised of a catalytic ATPase subunit, SNF2h, and an accessory subunit, Acf1(also known as SMARCA5 and BAZ1A, respectively)[19–22], was added to the surface-anchored nucleosomes to induce remodelling. FRET decreased substantially on addition of ACF and ATP (Fig. 1b), whereas incubation with ACF alone resulted in no significant change in FRET (data not shown). The observed decrease in FRET is consistent with the ability of ACF to centre mononucleosomes on DNA[10,11,23,26,27] (Fig. 1a). The average remodelling rate measured from nucleosomes anchored to the surface was quantitatively similar to that determined from measurements of nucleosomes in solution, indicating that surface-anchoring of nucleo-somes did not inhibit the activity of ACF (Supplementary Fig. 3).

In order to correlate the observed FRET value to the octamer posi-tion quantitatively, we measured FRET for a series of nucleosome constructs with different linker DNA lengths (n) on the exit side (Fig. 1c). The FRET value decreased monotonically with increasing exit linker length in a manner similar to the distance-dependence of FRET observed between donor and acceptor dyes attached to a DNA duplex (Supplementary Fig. 4). To test further whether the ACF-induced
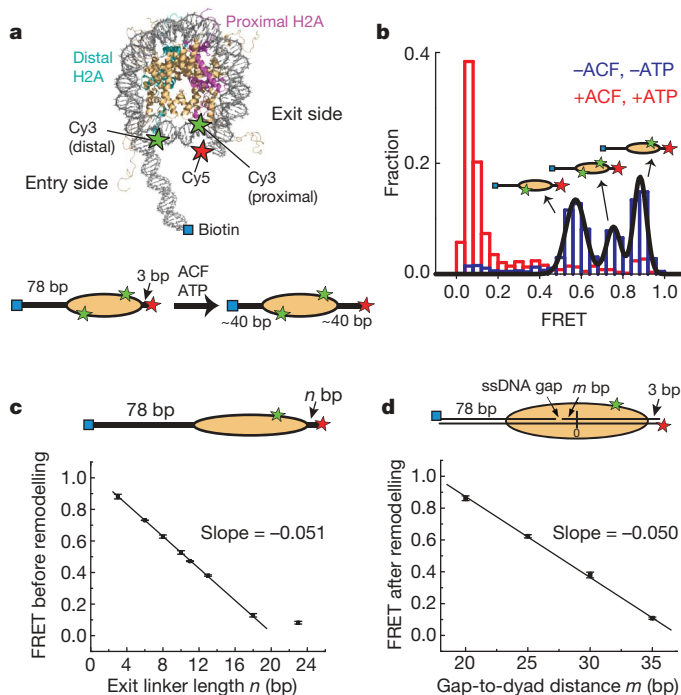
Figure 1 | Monitoring ACF-catalysed nucleosome remodelling by single-molecule FRET. a, The nucleosome structure[33] is shown (upper panel) with labelling sites for Cy3 and Cy5 indicated by green and red stars, respectively. Additional B-form DNA is modelled onto the entry and exit sides of the nucleosome to show the flanking DNA linkers. A linear nucleosome scheme (lower panel) showing the footprint of the histone octamer (yellow oval) on the DNA (black line) before and after ACF-catalysed remodelling. **b**, The FRET distribution of the $n = 3$ bp nucleosomes before (blue bars) and after (red bars) remodelling. The three initial peaks centred at FRET = 0.88, 0.75 and 0.58 (derived from Gaussian fit, black line) result from the three distinct Cy3-labelling configurations. After equilibration with ACF and ATP, the FRET values reduce to below 0.1. **c**, The initial FRET value as a function of the exit linker DNA length (n). The data were fit to a line with a slope of -0.051 ± 0.002 (black line). The last point near zero FRET is excluded from the linear fit. In this and subsequent figures, data from nucleosomes with a single Cy3 dye on the proximal H2A subunit are presented. The selection criteria for these nucleosomes are described in Methods. **d**, The final FRET values after remodelling by ACF as a function of m, the number of base pairs between the ssDNA gap and the nucleosome dyad (denoted as 0). The linear fit (black line) gives a slope of -0.050 ± 0.002. Error bars are $\pm$ s.e.m.

FRET change was indeed due to translocation of the histone octamer on DNA, we designed nucleosomes with stall sites defined by single-stranded (ss) DNA gaps. It has been shown that the ATPase domain of ISWI remodellers contacts a DNA region two helical turns ($\sim$20 bp) from the dyad axis of the nucleosome, and that ssDNA gaps located in this region inhibit nucleosome translocation[28–30]. We thus prepared a series of nucleosomes with the same linker DNA lengths (78 bp on the entry side and 3 bp on the exit side), each possessing a two-nucleotide ssDNA gap at a specified distance (m bp) away from the dyad axis (Fig. 1d). Although the initial FRET values of these constructs were similar to that observed for the construct without the ssDNA gap, the final FRET values after remodelling showed a strong dependence on the position of the ssDNA gap (Fig. 1d), with little FRET change for the construct with $m = 20$ bp and a final FRET versus m slope identical to that observed for the exit linker length dependence shown in Fig. 1c. These results demonstrate that the observed ACF-induced FRET changes can be quantitatively interpreted in terms of nucleosome translocation along DNA, although we cannot formally exclude the possibility that other alterations in nucleosome structure could also make a minor contribution. An interesting consideration is the spontaneous site exposure owing to fraying DNA ends previously reported to occur in the 0.01–0.05 s time scale[31], which should not cause significant fluctuations in FRET observed here with 0.1–2 s time resolution.

Multiple ATP-dependent remodelling steps

Next we characterized the remodelling kinetics by adding ACF and ATP to the nucleosomes *in situ* during data acquisition. After the addition of ACF and ATP, individual nucleosomes exhibited a 'waiting' period before any detectable change in FRET, followed by a 'translocation' period, during which FRET decreased to the background level (Fig. 2a). The duration of the waiting period (t_{wait}) depended on both ACF and ATP concentrations (Fig. 2b). The distributions of t_{wait} obtained at various ACF and ATP conditions indicate that the waiting phase included at least two steps, one depending on the ACF concentration and the other on ATP (Supplementary Fig. 5). To determine the order of these two steps, we performed a three-colour experiment with dye-labelled ACF, in which signal from the Alexa 488 dye on ACF directly reported the binding of the enzyme, whereas the FRET pair on the nucleosome reported the nucleosome position on the DNA. Notably, the binding of ACF preceded the onset of FRET decrease (Fig. 2c). Both the time before ACF binding (t_{bind}) and the time lag (t_{lag}) from ACF binding to the onset of FRET decrease depended on the ATP concentration (Fig. 2c), indicating that the waiting phase consisted of an ATP-dependent ACF binding step followed by an additional ATP-dependent step after the enzyme bound.

In contrast to the waiting phase, the duration of the translocation phase ($t_{translocate}$) was only dependent on ATP but not on ACF concentration (Fig. 2d and Supplementary Fig. 6), indicating that binding of additional ACF molecules was not required during this phase. Consistent with this notion, when we prebound nucleosomes with ACF and then removed unbound ACF with a buffer containing ATP but not ACF to initiate remodelling, the majority (86%) of the remodelled nucleosomes showed a complete decrease in FRET to below 0.1, indicating that the translocation phase did not require binding of additional ACF from the solution.

Translocation pauses during remodelling

Notably, translocation of the nucleosome did not proceed at a constant rate. Instead, the translocation phase exhibited periods of gradual decrease in FRET interrupted by translocation pauses (Fig. 3). For nucleosomes with the initial exit linker length $n = 3$ bp, the first pause occurred at a FRET value of 0.53 ± 0.03 (Fig. 3a, b), corresponding to an increase of linker length to 9.9 ± 0.6 bp and thus nucleosome translocation by 6.9 ± 0.6 bp. The pause position seemed to be independent of the initial linker length: for nucleosome constructs with four different linker DNA lengths ($n = -3, 0, 3$ and 6 bp), the first pause all occurred after approximately 7 bp of DNA translocation (Supplementary Fig. 7a).

In addition, we tested the dependence of the pause position on DNA sequence using a weaker positioning sequence 'A-100' (Supplementary Fig. 1b), which has $\sim$100 fold lower affinity than the 601 sequence[32]. The first pause of these nucleosomes again occurred after approximately 7 bp of translocation (Supplementary Fig. 7b). Although we cannot formally rule out the possibility that the positioning sequences contributed to the position of this initial pause, the observation that nucleosomes with two substantially different DNA sequences exhibit the same initial pause position indicates a potentially general feature of remodelling by ACF.

In addition to the first pause, subsequent translocation pauses were observed at lower FRET values (Fig. 3a, c). For nucleosomes with initial exit linker length $n = 3$ bp, a second and third pause preferentially occurred at FRET = 0.34 ± 0.03 and 0.17 ± 0.03, corresponding to 3.8 ± 0.6 bp and 3.3 ± 0.6 bp of translocation before pausing, respectively (Fig. 3a, b). Similar pauses were also observed for the $n = -3$ bp nucleosomes, except that the shorter exit linker length after the third pause allowed detection of a fourth pause, which occurred after 3.6 ± 0.8 bp of translocation from the third pause (Figs 3c, d). Taken together, these results indicate that the nucleosomes were translocated by a shorter distance (3–4 bp) between the subsequent pauses. Both the dwell time of the pauses and the duration of the translocation phases in between pauses
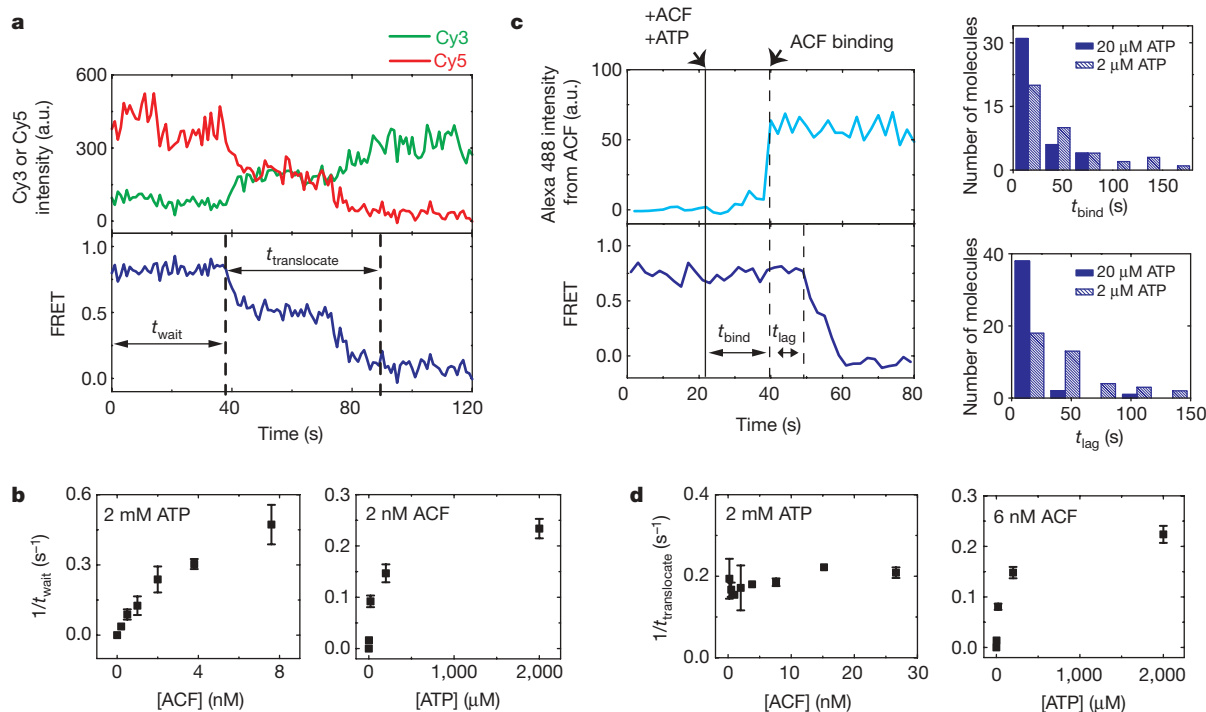
1023

Figure 2 | Real-time dynamics of ACF-catalysed nucleosome translocation.
a, Donor fluorescence (green), acceptor fluorescence (red) and FRET (blue) traces showing the ACF-induced remodelling of a single nucleosome ($n = 3$ bp). ACF (6 nM) and ATP (2 μM) were added at time zero. The durations of the waiting phase and the translocation phase are denoted as t_{wait} and $t_{translocate}$, respectively. a.u., Arbitrary units. **b**, Dependence of the mean t_{wait} value on ACF and ATP concentrations. **c**, Simultaneous monitoring of the binding of ACF and the remodelling of nucleosomes. Left panels: the upper trace shows the fluorescence signal from the Alexa 488-labelled ACF; the lower trace shows FRET between Cy3 and Cy5 on the nucleosome. ACF (4 nM) and ATP (20 μM) were added at the time indicated by the solid black line. The binding event of ACF (indicated by the first dashed line) further divides t_{wait} into two phases, t_{bind} and t_{lag}. Right panels: the distributions of t_{bind} and t_{lag} at two different ATP concentrations and 4 nM ACF. The distributions at different ATP concentrations are statistically distinct with 95% confidence for t_{bind} and more than 99% confidence for t_{lag} according to the Kolmogorov–Smirnov test. **d**, Dependence of the mean $t_{translocate}$ value on ACF and ATP concentrations. Error bars are $\pm$ s.e.m.

depended on the concentration of ATP, indicating that ATP binding was required in both phases (Fig. 3e and Supplementary Fig. 8). The dwell times of the subsequent pauses were similar to each other but substantially shorter than that of the first pause.

We note that the sum of a 7 bp and a 3–4 bp step and the sum of three 3–4 bp steps are both close to the 10 bp periodicity of DNA–histone contacts within the nucleosome[33]. Interestingly, the remodelling intermediates at a fraction of the periodicity (7 bp and 3–4 bp) were not stable in the absence of the remodelling enzyme: after removal of ACF, these intermediates collapsed to nucleosomal states in which the histone octamer was repositioned by a multiple of ~10 bp from the pre-remodelling position (Supplementary Fig. 9). These collapsed states, consistent with the previously observed accumulation of remodelling products at ~10 bp intervals of nucleosome translocation[28,29], are probably imposed by structural constraints of the nucleosome.

Processive and bidirectional translocation

The above experiments with end-positioned nucleosomes provide quantitative analyses of remodelling kinetics and intermediates. The limited dynamic range of FRET, however, made it difficult to characterize the equilibrium state(s) after remodelling using these substrates. Considering that ACF tends to centre the nucleosome on the DNA, we reasoned that a centre-positioned nucleosome with an initial FRET value within the dynamic range of FRET would facilitate the analysis of equilibrium remodelling dynamics. To this end we constructed a centre-positioned mononucleosome with the 601 sequence flanked by 78 bp of DNA on each side and an internal acceptor label (Fig. 4a, Supplementary Fig. 1a). The initial FRET distribution showed a narrow peak at FRET = 0.3 (Supplementary Fig. 10). After equilibration with ACF and ATP, the FRET distribution broadened substantially (Supplementary Fig. 10) and the time traces of individual nucleosomes exhibited large-amplitude oscillations in FRET

(Fig. 4b), indicating that the histone octamer was translocated back-and-forth along the DNA by the remodelling enzyme. Bidirectional remodelling was observed to be the predominant behaviour (>70% of remodelled nucleosomes), even at sub-saturating conditions in which a low concentration (1 nM) of ACF was added to induce remodelling of only a small fraction (<10%) of the nucleosomes. Autocorrelation analysis of these FRET time traces showed a characteristic oscillation time that was dependent on the ATP concentration but independent of the ACF concentration (Fig. 4b), indicating that the observed bidirectional translocation was accomplished by continuously bound ACF without requiring dissociation and rebinding of ACF from the solution. To test further this notion, we performed three-colour experiments with Alexa 488-labelled ACF and FRET-labelled nucleosomes, in which signal from the Alexa 488 dye directly reported the binding of ACF. Repeated back-and-forth movement of the nucleosomes was observed within individual ACF binding events (Supplementary Fig. 11), further confirming that the bidirectional nucleosome translocation was accomplished by a continuously bound ACF complex.

To quantify further the processivity of ACF, we performed buffer exchange experiments in which ACF and ATP were added and unbound ACF (but not ATP) was subsequently removed *in situ* as the position of individual nucleosomes was monitored. Notably, the ACF-induced bidirectional movement persisted for a long period of time after unbound ACF was removed from the solution (Fig. 4c). The nucleosomes were translocated with an average speed of approximately 2 bp s⁻¹. The lower-bound estimate of the cumulative distance travelled by the nucleosome after removal of unbound ACF exhibits a broad distribution with a mean of 200 bp (Fig. 4c). Taken together, these results indicate that ACF is a highly processive and bidirectional nucleosome translocase. The observed processivity is consistent with the strong commitment of ISWI enzymes to nucleosomal templates once chromatin assembly and remodelling are initiated[34,35].
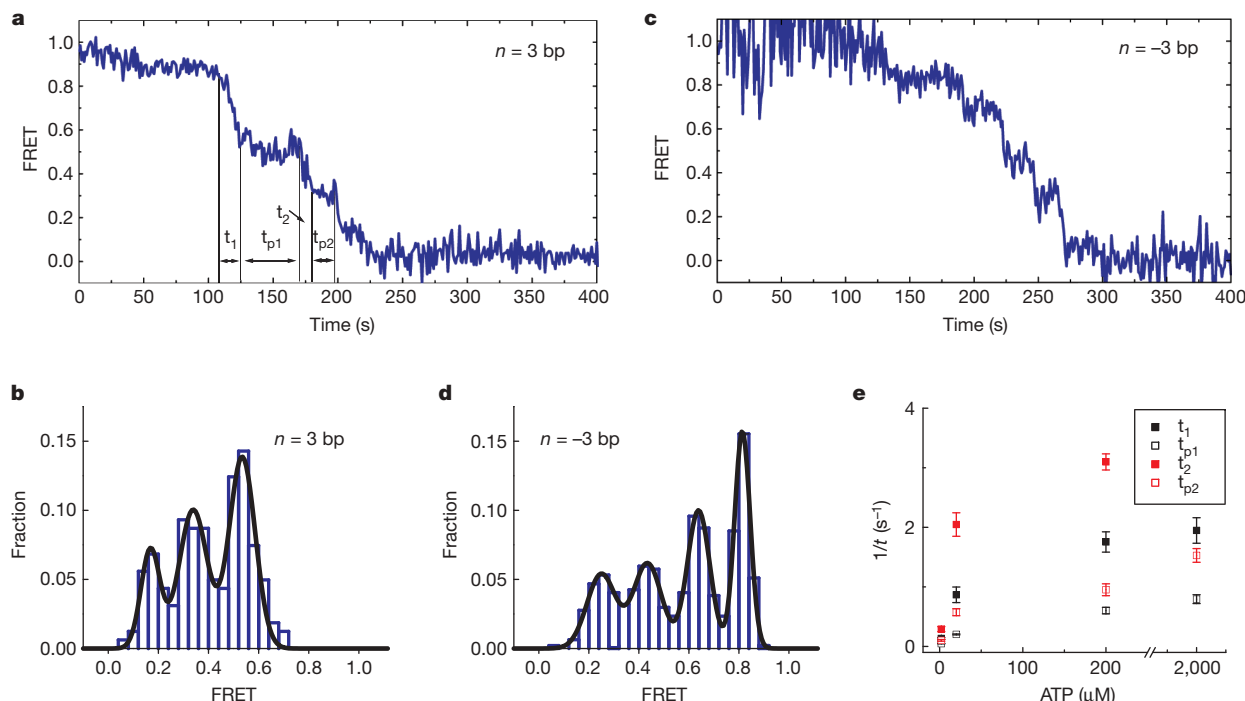
Figure 3 | ACF-catalysed nucleosome translocation is interrupted by well defined kinetic pauses. a, FRET time trace of a nucleosome ($n = 3$ bp) showing kinetic pauses that divide the entire translocation phase into several translocation and pause sub-phases: t_1, t_{p1}, t_2, t_{p2}... ACF (6 nM) and ATP (2 μM) were added at time zero. **b,** FRET distribution of the pauses constructed from many $n = 3$ bp nucleosomes. The peak FRET values, 0.53, 0.34, and 0.17 (obtained from Gaussian fit, black line), correspond to 6.9 bp of translocation between the initial position and the first pause, 3.8 bp between the first and second pauses, and 3.3 bp between the second and third pauses.

c, FRET time trace of a $n = -3$ bp nucleosome after addition of ACF (6 nM) and ATP (2 μM) at time zero. The proximity of the initial donor and acceptor positions causes partial quenching of their fluorescence and thus relatively large fluctuations in initial FRET. **d,** FRET distribution of the pauses constructed from many $n = -3$ bp nucleosomes. The peaks correspond to 7.3 bp of translocation between the initial position and the first pause, 3.4 bp between the first and second pauses, 4.0 bp between the second and third pauses, and 3.6 bp between the third and fourth pauses. **e,** ATP-dependence of the mean t_1, t_{p1}, t_2, and t_{p2} values. Error bars are ± s.e.m.

It is striking that an ACF complex remaining bound to the nucleosome could cause such a highly processive, back-and-forth nucleosome movement. Such a demanding task could be accomplished if ACF preferentially binds the nucleosome as a dimer, in which two ACF monomers, particularly their corresponding ATPase domains, are bound on opposite sides of the nucleosome and oriented for translocation in opposing directions. Coordinated action of the two monomers would then allow processive back-and-forth translocation of the nucleosome. This hypothesis is supported by our three-colour experiments with Alexa 488-labelled ACF and FRET-labelled nucleosomes. To determine the number of ACF bound to the nucleosome, we performed statistical analyses of the Alexa 488 intensity and the number of Alexa 488 photobleaching steps associated with each ACF binding event. These analyses are statistically consistent with the notion that the binding events leading to bidirectional nucleosome remodelling contained two ACF monomers, whereas the binding events leading to unidirectional remodelling contained a single ACF monomer (Supplementary Fig. 12). Further supporting this model, electron microscopy and biochemical data showed cooperative binding of two SNF2h proteins to a single nucleosome with each SNF2h occupying one side of the nucleosome in an activated ATP state[36]. The diffusion coefficient of ACF bound to DNA is also consistent with a complex of two Acf1 and two SNF2h subunits[37]. Interestingly, the SWI/SNF subfamily enzymes can also reversibly create and retract DNA loops[12,13], but it is unclear whether the bidirectional nucleosome translocation by ACF and the reversible DNA loop formation by SWI/SNF share a common mechanism.

Discussion

We have developed a single-molecule assay to monitor the remodelling of individual nucleosomes by chromatin remodelling enzymes in

real time. This assay allowed quantitative characterization of the structural dynamics and kinetic intermediates of nucleosomes during remodelling. Using this approach, we showed that the human ACF enzyme induced gradual translocation of nucleosomes along DNA interrupted by well-defined kinetic pauses. ATP has multiple functional roles in the remodelling process. The three distinct steps during remodelling, namely binding of ACF, translocation of the nucleosome and translocation pauses, were all ATP-dependent, demonstrating a versatile usage of ATP by an enzyme with only one type of ATP binding site.

Quantification of the FRET traces of end-positioned nucleosomes showed that the first kinetic pause occurred after approximately 7 bp of nucleosome translocation, whereas subsequent pauses were separated by only 3–4 bp. Although it is currently unclear whether these remodelling intermediates occur only at the beginning of remodelling or continue into the processive remodelling phase, similar translocation pauses were also observed during the continuous remodelling process of centre-positioned nucleosomes (Fig. 4b) and thus may represent a fundamental property of ACF-induced remodelling. One possible origin of these intermediates is an ATP-dependent conformational change of the remodelling enzyme that prepares the nucleosome for the next round of DNA translocation (for example, by forming a DNA loop for subsequent propagation around the nucleosome)[30,38,39]. The unique properties of the first pause, compared to the subsequent pauses, may indicate a complex initiation phase of remodelling.

On centre-positioned nucleosomes, ACF was observed to exhibit remarkable processivity and bidirectionality: an ACF complex continuously bound to a nucleosome could translocate the histone octamer back-and-forth by a total distance of more than 200 bp and switch directions more than 20 times on average before dissociation.
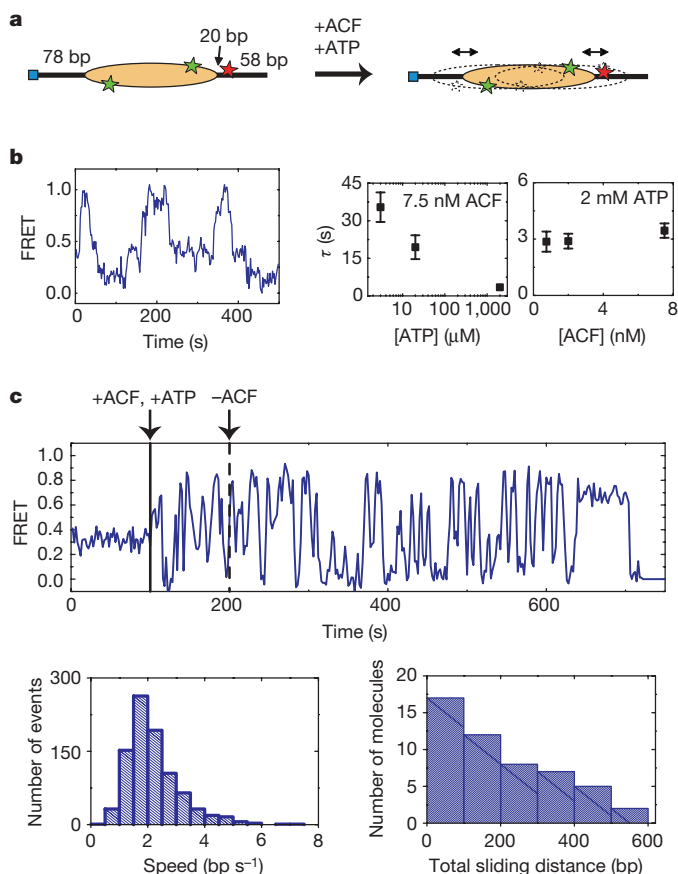
a

b

c

Figure 4 | ACF catalyses processive and bidirectional nucleosome translocation. a, A centre-positioned nucleosome flanked by 78 bp of linker DNA on both sides is subject to remodelling. **b,** FRET trace of a nucleosome in equilibrium with 7.5 nM ACF and 3 μM ATP showing back-and-forth translocation on DNA(left panel); the characteristic time of the FRET fluctuations (τ) depends on ATP, but not on ACF, concentration (right panel). The τ values were derived from autocorrelation analysis as described in Methods. **c,** Upper panel, a FRET trace showing processive and bidirectional nucleosome translocation by a continuously bound ACF complex. ACF (3 nM) and ATP (2 mM) were added at the time indicated by the solid black line. Unbound ACF (but not ATP) was then removed from the solution at the time indicated by the dashed black line. Lower left panel distribution of the translocation speed within each segment of unidirectional translocation. Lower right panel, distribution of the cumulative distance travelled by individual nucleosomes after removal of unbound ACF. Estimate of the travelling speed and distance is described in Methods. Error bars are ± s.e.m.

Statistical analyses indicate that the bidirectional remodelling is most probably caused by ACF dimers. The processive and bidirectional translocation of nucleosomes potentially allows ACF to rapidly sample the DNA on both sides of the nucleosome to generate regular inter-nucleosomal spacing.

METHODS SUMMARY

Briefly, various mononucleosome constructs, with different DNA sequences, DNA linker lengths and ssDNA gap locations, were reconstituted using histone octamers that were labelled with a green FRET donor dye (Cy3) and double stranded DNA that was labelled with a red FRET acceptor dye (Cy5) and a biotin. The nucleosomes were then anchored to a microscope slide via a biotin–streptavidin linkage. Unlabelled ACF or ACF labelled with a blue dye (Alexa 488) were added to the surface-anchored nucleosomes together with ATP to induce remodelling. The fluorescence signals from Alexa 488, Cy3 and Cy5 were detected by a TIRF microscope, separated by dichroic mirrors, and imaged onto separate areas of an amplified-CCD camera after passing through various fluorescence emission filters. Custom-written software was used to identify single nucleosomes on the slide and to monitor the Alexa 488, Cy3 and Cy5 fluorescence at these positions for extended periods of time. The FRET value was defined as $I_A/(I_D + I_A)$, where I_D

and I_A represent the fluorescence signals detected in the Cy3 and Cy5 channels, respectively.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Becker, P. B. & Horz, W. ATP-dependent nucleosome remodeling. *Annu. Rev. Biochem.* **71**, 247–273 (2002).
2. Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
3. Flaus, A. & Owen-Hughes, T. Mechanisms for ATP-dependent chromatin remodelling: farewell to the tuna-can octamer? *Curr. Opin. Genet. Dev.* **14**, 165–173 (2004).
4. Smith, C. L. & Peterson, C. L. ATP-dependent chromatin remodeling. *Curr. Top. Dev. Biol.* **65**, 115–148 (2004).
5. Clapier, C. R. & Cairns, B. R. The biology of chromatin remodeling complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
6. Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodeling by RSC involves ATP-dependent DNA translocation. *Genes Dev.* **16**, 2120–2134 (2002).
7. Whitehouse, I., Stockdale, C., Flaus, A., Szczelkun, M. D. & Owen-Hughes, T. Evidence for DNA translocation by the ISWI chromatin-remodeling enzyme. *Mol. Cell. Biol.* **23**, 1935–1945 (2003).
8. Tsukiyama, T., Palmer, J., Landel, C. C., Shiloach, J. & Wu, C. Characterization of the Imitation Switch subfamily of ATP-dependent chromatin-remodeling factors in *Saccharomyces cerevisiae. Genes Dev.* **13**, 686–697 (1999).
9. Hamiche, A., Sandaltzopoulos, R., Gdula, D. A. & Wu, C. ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF. *Cell* **97**, 833–842 (1999).
10. Längst, G., Bonte, E. J., Corona, D. F. & Becker, P. B. Nucleosome movement by CHRAC and ISWI without disruption or trans-displacement of the histone octamer. *Cell* **97**, 843–852 (1999).
11. Kassabov, S. R., Henry, N. M., Zofall, M., Tsukiyama, T. & Bartholomew, B. High-resolution mapping of changes in histone-DNA contacts of nucleosomes remodeled by ISW2. *Mol. Cell. Biol.* **22**, 7524–7534 (2002).
12. Zhang, Y. *et al.* DNA translocation and loop formation mechanism of chromatin remodeling by SWI/SNF and RSC. *Mol. Cell* **24**, 559–568 (2006).
13. Lia, G. *et al.* Direct observation of DNA distortion by the RSC complex. *Mol. Cell* **21**, 417–425 (2006).
14. Shundrovsky, A., Smith, C. L., Lis, J. T., Peterson, C. L. & Wang, M. D. Probing SWI/SNF remodeling of the nucleosome by unzipping single DNA molecules. *Nature Struct. Mol. Biol.* **13**, 549–554 (2006).
15. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc. Natl Acad. Sci. USA* **58**, 719–726 (1967).
16. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* **93**, 6264–6268 (1996).
17. Zhuang, X. *et al.* A single molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051 (2000).
18. Ito, T., Bulger, M., Pazin, M. J., Kobayashi, R. & Kadonaga, J. T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145–155 (1997).
19. Ito, T. *et al.* ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. *Genes Dev.* **13**, 1529–1539 (1999).
20. Bochar, D. A. *et al.* A family of chromatin remodeling factors related to Williams syndrome transcription factor. *Proc. Natl Acad. Sci. USA* **97**, 1038–1043 (2000).
21. LeRoy, G., Loyola, A., Lane, W. S. & Reinberg, D. Purification and characterization of a human factor that assembles chromatin. *J. Biol. Chem.* **275**, 14787–14790 (2000).
22. Poot, R. A. *et al.* HuCHRAC, a human ISWI chromatin remodelling complex contains hACF1 and two novel histone-fold proteins. *EMBO J.* **19**, 3377–3387 (2000).
23. Yang, J. G., Madrid, T. S., Sevastopoulos, E. & Narlikar, G. J. The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. *Nature Struct. Mol. Biol.* **13**, 1078–1083 (2006).
24. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
25. Abbondanzieri, E. A. *et al.* Dynamic binding orientations direct activity of HIV reverse transcriptase. *Nature* **453**, 184–189 (2008).
26. He, X., Fan, H. Y., Narlikar, G. J. & Kingston, R. E. Human ACF1 alters the remodeling strategy of SNF2h. *J. Biol. Chem.* **281**, 28636–28647 (2006).
27. Stockdale, C., Flaus, A., Ferreira, H. & Owen-Hughes, T. Analysis of nucleosome repositioning by yeast ISWI and Chd1 chromatin remodeling complexes. *J. Biol. Chem.* **281**, 16279–16288 (2006).
28. Schwanbeck, R., Xiao, H. & Wu, C. Spatial contacts and nucleosome step movements induced by the NURF chromatin remodeling complex. *J. Biol. Chem.* **279**, 39933–39941 (2004).

29. Zofall, M., Persinger, J., Kassabov, S. R. & Bartholomew, B. Chromatin remodeling by ISW2 and SWI/SNF requires DNA translocation inside the nucleosome. *Nature Struct. Mol. Biol.* **13**, 339–346 (2006).

30. Dang, W. & Bartholomew, B. Domain architecture of the catalytic subunit in the ISW2-nucleosome complex. *Mol. Cell. Biol.* **27**, 8306–8317 (2007).

31. Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nature Struct. Mol. Biol.* **12**, 46–53 (2004).

32. Partensky, P. D. & Narlikar, G. J. Chromatin remodelers act globally, sequence positions nucleosomes locally. *J. Mol. Biol.* **391**, 12–25 (2009).

33. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

34. Fyodorov, D. V. & Kadonaga, J. T. Dynamics of ATP-dependent chromatin assembly by ACF. *Nature* **418**, 896–900 (2002).

35. Gangaraju, V. K., Prasad, P., Srour, A., Kagalwala, M. N. & Bartholomew, B. Conformational changes associated with template commitment in ATP-dependent chromatin remodeling by ISW2. *Mol. Cell* **35**, 58–69 (2009).

36. Raki, L. R. The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. *Nature* doi:10.1038/nature08621 (this issue).

37. Strohner, R. *et al.* A 'loop recapture' mechanism for ACF-dependent nucleosome remodeling. *Nature Struct. Mol. Biol.* **12**, 683–690 (2005).

38. Fitzgerald, D. J. *et al.* Reaction cycle of the yeast Isw2 chromatin remodeling complex. *EMBO J.* **23**, 3836–3843 (2004).

39. Cairns, B. R. Chromatin remodeling: insights and intrigue from single-molecule studies. *Nature Struct. Mol. Biol.* **14**, 989–996 (2007).

Author Contributions T.R.B. performed the experiments and analysis with help from M.D.S.; J.G.Y. made the enzymes and histone proteins. T.R.B., G.J.N. and X.Z. designed the experiments. X.Z. oversaw the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to X.Z. (zhuang@chemistry.harvard.edu) and G.J.N (geeta.narlikar@ucsf.edu).

METHODS

Preparation of mononucleosomes. Various double-stranded (ds) DNA constructs with different sequences and flanking linker lengths, as shown in Supplementary Fig. 1, were made by PCR and purified by PAGE. When needed, two-nucleotide (nt) single-stranded DNA (ssDNA) gaps were generated in the dsDNA by excision of two adjacent deoxyuridine residues by the Uracil-Specific Excision Reagent (USER, New England Biolabs), which contains a mixture of uracil DNA glycosylase and endonuclease VIII[29]. Before USER digestion, the full-length dsDNA was constructed by ligation of two separate, PAGE-purified, PCR products with complementary sticky ends made by restriction digestion. Deoxyuridine, 5′-Cy5 and 5′-biotin modifications were incorporated synthetically into PCR primers (Integrated DNA Technologies). For internal Cy5 labelling, PCR primers containing an internal amino modifier (dT C6) (Integrated DNA Technologies) were labelled with a monoreactive Cy5 (GE Healthcare) and purified by reverse-phase HPLC over a C8 column (GE Healthcare) before use in PCR.

To label the histone H2A protein with the FRET donor, a unique cysteine substitution was introduced to residue 120 of H2A. A Cy3 dye (GE Healthcare) was attached to the cysteine as described previously[23]. The Cy3-labelled and unlabelled H2A were mixed at a ratio of 1:1 together with other histone proteins (H2B, H3 and H4) to form histone octamers. Mononucleosomes were reconstituted from DNA and octamers by salt gradient dialysis and purified by ultracentrifugation over a 10–30% (v/v) glycerol gradient[40]. Without addition of the remodelling enzyme, the FRET values from the nucleosomes remained unchanged for at least one hour after the nucleosomes were anchored to the surface, indicating that the nucleosomes were stable during this time scale.

Preparation of ACF. The procedure to prepare ACF was previously described[23,41]. Briefly, haemagglutinin (HA)–Acf1 and Flag–SNF2h were individually overexpressed in Sf9 cells using a baculovirus expression system. Excess HA-tagged Acf1 extract was mixed with Flag-tagged SNF2h extract, and the ACF complex was purified using M2-affinity chromatography[41]. For experiments that directly monitor the binding of ACF to nucleosomes, we labelled the ACF complex with the Alexa 488 dye. Before M2 elution, the resin was washed with labelling buffer (20% glycerol, 20 mM HEPES pH 7.0, 0.2 mM EDTA, 100 mM KCl, 1 mM benzamidine HCl, 1 mM TCEP) followed by addition of Alexa 488 maleimide (Molecular Probes) to 100 µM final concentration. After 30 min at 4 °C, the labelling reaction was quenched with 80 mM β-mercaptoethanol for 15 min. The resin was washed extensively with wash buffer (20% glycerol, 20 mM HEPES pH 7.9, 0.2 mM EDTA, 100 mM KCl, 1 mM benzamidine HCl, 1 mM DTT) and the protein was then eluted. Stoichiometry of Acf1 to SNF2h was confirmed by SYPRO staining.

Single-nucleosome FRET measurements. Quartz microscope slides were coated with methoxy-poly(ethylene glycol) (PEG, Nektar Therapeutics or Laysan Bio), biotin-PEG (Nektar Therapeutics or Laysan Bio) and streptavidin as described previously[25]. The biotinylated nucleosomes were then linked to the streptavidin-coated slide surface via a biotin-streptavidin linkage. The donor and acceptor fluorescence signals from the surface-anchored nucleosomes were excited with a 532 nm Nd:YAG (neodymium-doped yttrium aluminium garnet) laser (CrystaLaser) in the total internal refection geometry and fluorescence emission from Cy3 and Cy5 was detected with a ×60 water immersion objective (Olympus), filtered with a 550 nm long-pass filter (Chroma Technology), spectrally spilt by a 630 nm dichroic mirror (Chroma Technology), and imaged onto two halves of a CCD camera (Andor iXon 897 or iXon^{EM}+ 888; ref. 25). Unlabelled ACF was added to the surface-anchored nucleosomes together with ATP to induce remodelling. In experiments that monitored both the binding of ACF and the remodelling of nucleosomes, ACF labelled with Alexa 488 was added to the nucleosomes. Alexa 488 and Cy3 were excited by alternating 488 nm argon ion and 532 nm Nd:YAG laser lines, respectively.

Because the FRET donor Cy3 is attached to the H2A subunit of the histone octamer, the presence of two H2A subunits in each octamer led to a heterogeneous population of nucleosomes with three different labelling configurations: (1) donor on the H2A subunit proximal to the acceptor, yielding a high FRET level, (2) donor on the distal H2A, yielding a lower FRET level, (3) donor on both H2A subunits, yielding an intermediate FRET value. For example, three distinct peaks centred at FRET = 0.88, 0.75 and 0.58 were observed in the FRET distribution of the nucleosomes with exit linker length $n = 3$ bp (Fig. 1b). In this work, we focus our analyses on nucleosomes containing a single donor on the proximal H2A. In the FRET histogram analysis, the histograms were fit with three Gaussian peaks and the peak with the highest FRET value was selected. In the FRET trace analysis, traces starting with a mean FRET > 0.75 and exhibiting a single donor bleaching step were selected. This selection process allowed us to include the entire population of nucleosomes with a donor on the proximal H2A without contamination from nucleosomes with donor on both H2A subunits, which exhibit two donor

bleaching steps. Such selection process indeed resulted in a single population of nucleosomes with FRET centred at 0.88 (Supplementary Fig. 2e). Similar selection criteria were used for nucleosomes with exit linker length $n = 6$ bp, 0 bp and -3 bp: Traces starting with a mean FRET larger than a threshold value (0.6 for $n = 6$ bp, 0.75 for $n = 0$ bp and $n = -3$ bp) and exhibiting a single donor bleaching step were selected for further analysis. To generate the calibration curve in Fig. 1c, nucleosomes with $n = 3, 6, 8, 10, 11, 13, 18$ and 23 bp were used. The FRET distributions in each case were fit to three Gaussians and the peak with the highest FRET value was selected, except for the $n = 18$ or 23 bp nucleosomes, which exhibited only very low FRET values that appeared as one peak. In these two latter cases, the FRET distribution was fit to a single Gaussian function to extract the peak position.

In experiments where buffer exchange was used to add or remove ACF and ATP, the sample chamber was infused with new buffer using a syringe pump (KD Scientific). To ensure complete buffer exchange within the sample chamber (~20 µl), a large excess of the new buffer was flown through the chamber on infusion. In the case of ACF and ATP addition, 300 µl or ~15 chamber volumes of new buffer was infused. In the case of ACF removal, 600 µl of new buffer was infused. The dead time for buffer exchange was measured to be 1.3 s.

Nucleosomes were imaged at 30 °C in a buffer consisting of 12 mM HEPES, 40 mM Tris pH 7.5, 60 mM KCl, 0.32 mM EDTA, 3 mM MgCl$_2$, 10% glycerol, 0.02% Igepal (Sigma Aldrich), an oxygen scavenging system (10% glucose, 800 µg ml^{-1} glucose oxidase, 40 µg ml^{-1} catalase) to reduce photobleaching, 2 mM Trolox (Sigma) to reduce photoblinking of the dyes[42], and 0.1 mg ml^{-1} BSA (Promega) to prevent non-specific sticking of nucleosomes and ACF to the surface.

Estimate of the characteristic oscillation time, the translocation speed and the total travelling distance of centre-positioned nucleosomes. The characteristic FRET fluctuation time (τ) of the centre-positioned nucleosomes in the presence of ACF and ATP was derived using autocorrelation analysis. Briefly, the autocorrelation function at each condition was constructed from ~100 FRET time traces. The characteristic time τ was derived as the decay constant from single exponential fitting of the autocorrelation function.

To estimate the cumulative distance (not net distance) of nucleosome translocation on DNA induced by a bound ACF complex as shown in Fig. 4c, FRET traces during the period after removal of ACF from the solution but before dissociation of the bound ACF from the nucleosome were divided into unidirectional segments of monotonically changing FRET. The FRET change in each segment was converted into the number of base pairs translocated by comparison with the calibration curve shown in Fig. 1c. The results from all segments of a trace were summed to give the total distance travelled by the octamer. The following reasons make the estimated total distance an approximate, lower bound estimate of the cumulative distance travelled by the histone octamer: (1) only unidirectional segments with a FRET change greater than 0.3 were selected for the sum to avoid counting noise in the FRET signal. To characterize noise, FRET fluctuations from the nucleosomes in the absence of ACF were analysed and the probability of observing a FRET change greater than 0.3 in the absence of ACF was found to be less than 0.002. Therefore, 0.3 is a rather conservative threshold for nucleosome translocation, leading to an underestimate of travelling distance. (2) The slope of the calibration curve in Fig. 1c was used to estimate distances from changes in FRET, but as the FRET value gets near 0 or 1, it saturates and becomes an insensitive measure of distance, also causing an underestimate. The possibility that the DNA trajectories within the nucleosome or on the entry side could differ from that of the exit linker DNA may give additional error in the distance estimate. To estimate the speed of octamer translocation, the distance change within each unidirectional segment determined above was divided by the duration of the segment. To determine the number of times a nucleosome switched its direction, only events switching from a decreasing to an increasing trend in FRET were counted, as the switching from an increasing trend to a decreasing trend could be due to the FRET acceptor on the DNA passing by the FRET donor on the octamer without switching direction. Therefore, the estimated number of switching times is also an underestimate, as some of the latter type of switching events may also represent real directional switching of the nucleosome translocation.

40. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol.* **304**, 3–19 (1999).
41. Aalfs, J. D., Narlikar, G. J. & Kingston, R. E. Functional differences between the human ATP-dependent nucleosome remodeling proteins BRG1 and SNF2H. *J. Biol. Chem.* **276**, 34270–34278 (2001).
42. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nature Methods* **3**, 891–893 (2006).

LETTERS

Two distinct sequences of blue straggler stars in the globular cluster M 30

F. R. Ferraro[1], G. Beccari[2], E. Dalessandro[1], B. Lanzoni[1], A. Sills[3], R. T. Rood[4], F. Fusi Pecci[5], A. I. Karakas[6], P. Miocchi[1] & S. Bovinelli[1]

Stars in globular clusters are generally believed to have all formed at the same time, early in the Galaxy's history[1]. 'Blue stragglers' are stars massive enough[2] that they should have evolved into white dwarfs long ago. Two possible mechanisms have been proposed for their formation: mass transfer between binary companions[3] and stellar mergers resulting from direct collisions between two stars[4]. Recently the binary explanation was claimed to be dominant[5]. Here we report that there are two distinct parallel sequences of blue stragglers in M 30. This globular cluster is thought to have undergone 'core collapse', during which both the collision rate and the mass transfer activity in binary systems would have been enhanced[6]. We suggest that the two observed sequences are a consequence of cluster core collapse, with the bluer population arising from direct stellar collisions and the redder one arising from the evolution of close binaries that are probably still experiencing an active phase of mass transfer.

To investigate the blue straggler star (BSS) content in M 30, we used a time series of 44 high-resolution images obtained with the NASA Hubble Space Telescope (Supplementary Information). The colour–magnitude diagram (CMD) derived by combining data from these images has revealed the existence of two well-separated and almost parallel sequences of BSSs (hereafter 'red BSSs' and 'blue BSSs'; Fig. 1). The two sequences are similarly populated, consisting of 21 and 24 stars, respectively.

The detected BSSs are substantially more concentrated towards the cluster centre than are 'normal' cluster stars, either along the subgiant branch or the horizontal branch (Fig. 2a). According to a Kolmogorov–Smirnov test, the probability that the BSSs and the subgiant- or horizontal-branch populations are drawn from the same distribution is only $\sim 10^{-3}$ (that is, they differ at a significance level of more than 4σ). This result confirms that BSSs are more massive[2] than the majority of the cluster stars and that mass segregation has been active in this cluster. Moreover, when we consider the distribution of the two BSS subpopulations separately, we find that the red BSSs are more centrally segregated than are the blue BSSs. Indeed, no red BSSs are observed at an angular distance of $r > 30$ arcsec (corresponding to ~ 1.3 pc; see Supplementary Table 1) from the cluster centre (Fig. 2a). Even though in this case the level of significance (1.5σ) is marginal because of the small number of objects, this evidence is suggestive of different formation histories for BSSs belonging to the two sequences. Furthermore, whereas the radial distribution of BSSs in many clusters is found to be bimodal[7] (with a dominant peak at the centre, a dip at intermediate radii and a rising branch in the outer regions), in the case of M 30 there is no evidence of an increase at large distances from the centre: more than 80% of the entire BSS population is confined within the inner 100 arcsec (~ 4.2 pc), and the radial distribution then is nearly constant
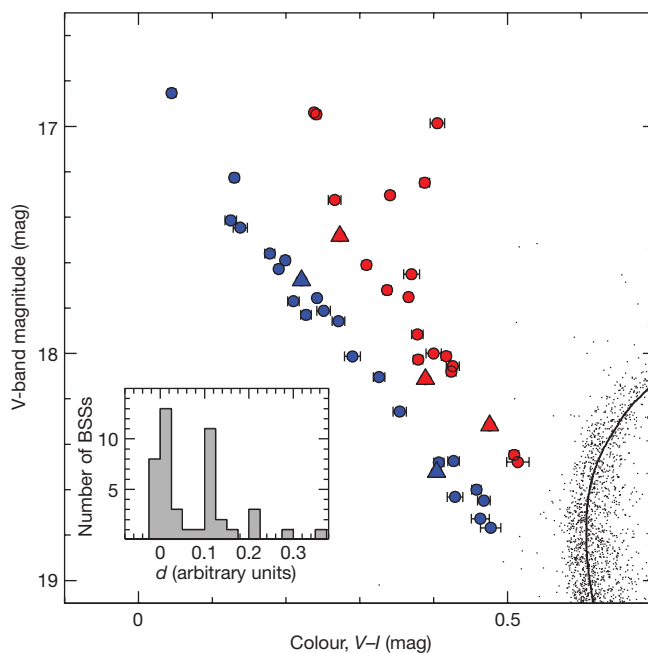

Figure 1 | The two blue straggler sequences of M 30. BSS region of the $(V, V-I)$ CMD. The selected BSSs are plotted as circles, with the red and blue colours distinguishing the red and the blue BSS sequences, respectively. A series of 44 images (22 in each filter), made using the Hubble Space Telescope's Wide Field Planetary Camera 2 (WFPC2) through the F814W (I band) and the F555W (V band) filters, has been analysed using point-spread-function-fitting photometry[21,22]. Errors (1 s.e.m.) in magnitude and colour have been computed from repeated measurements and are also plotted (they are typically less than 0.01 mag; in most cases the error bars are smaller than the point size). The two sequences are separated in magnitude by $\Delta V \approx 0.4$ mag and in colour by $\Delta(V-I) \approx 0.12$ mag. Using these time series, we have tested the variability of the selected BSSs and found five candidate variables (triangles): on the basis of the light-curve characteristics, the three brightest variables have been classified[23] as W Ursae Majoris (W UMa) contact binaries. The two faintest candidates have quite scattered light curves that prevent a reliable classification. W UMa stars are binary systems losing orbital momentum as a result of magnetic braking. These shrinking binary systems, which are initially detached, evolve to the contact stage and finally merge into a single star. The evolution of W UMa systems is thought to be a viable channel for the formation of BSSs[20,24]. Inset, distribution of the geometrical distances, d, of the selected BSSs from the straight line that best fits the blue BSS sequence. Two well-defined peaks are clearly visible. A dip test[25] applied to this distribution demonstrates that it is bimodal at a significance level of more than 4σ, confirming that the two sequences are nearly parallel to each other.

[1]Department of Astronomy, University of Bologna, Via Ranzani 1, 40127 Bologna, Italy. [2]ESA, Space Science Department, Keplerlaan 1, 2200 AG Noordwijk, Netherlands. [3]McMaster University, Department of Physics and Astronomy, 1280 Main Street West, Hamilton, Ontario L8S 4M1, Canada. [4]Astronomy Department, University of Virginia, PO Box 400325, Charlottesville, Virginia 22904, USA. [5]INAF, Osservatorio Astronomico di Bologna, Via Ranzani 1, 40127 Bologna, Italy. [6]Research School of Astronomy and Astrophysics, Mt Stromlo Observatory, Weston Creek, Australian Capital Territory 2611, Australia.
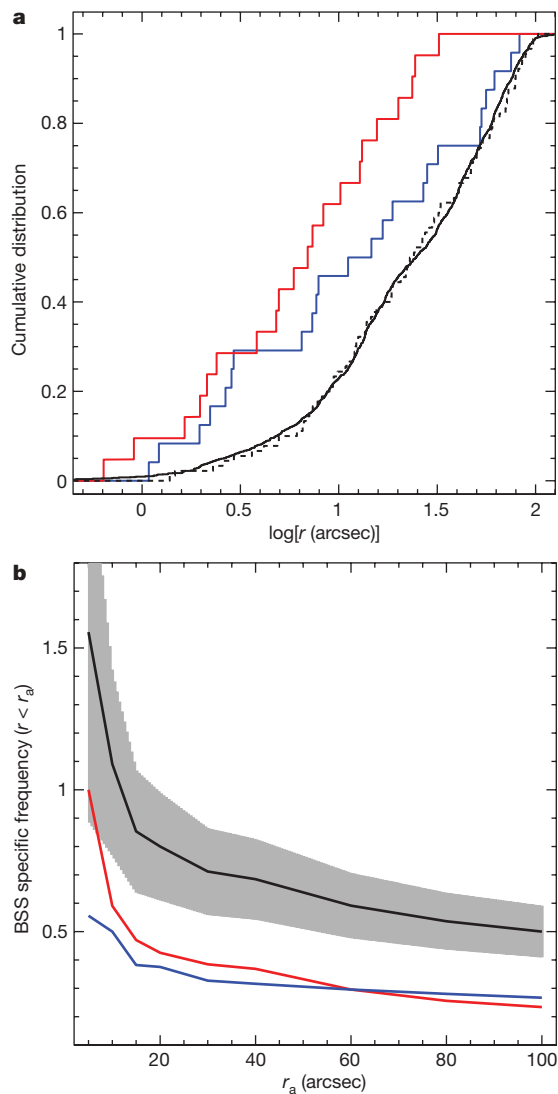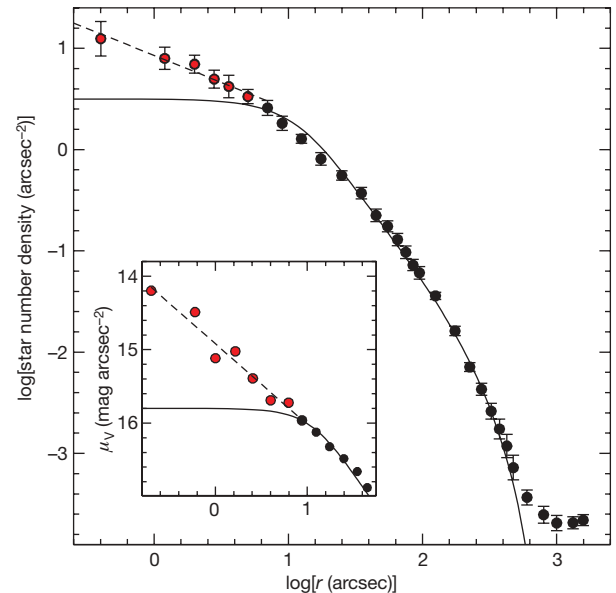
Figure 3 | The star density profile of M 30. Profile (errors, 1 s.e.m.) obtained from resolved star counts over the entire cluster extension: the WFPC2 data set was combined with Hubble Space Telescope Advanced Camera for Surveys data and with wide-field ground-based observations secured at the European Southern Observatory New Technology Telescope (La Silla Observatory, Chile) and the Canada–France–Hawaii Telescope MegaCam (Hawaii). The single-mass King model[26] that best fits the observed profile excluding the innermost ($r < 5$ arcsec) points is shown as a solid line. The points that deviate from the King model (red) are well fitted by a power law with slope $\alpha \approx -0.5$ (dashed line). Inset, surface brightness profile derived from the WFPC2 V-band images within the innermost 40 arcsec, with the two lines having the same meaning as above. The measured central surface brightness is $\mu_V \approx 14.2$ mag arcsec^{-2}. This value is significantly lower (corresponding to greater brightness) than that listed in currently used cluster catalogues[27], but is fully consistent with that obtained in most recent studies[28]. Following the procedure described in the literature[29] and assuming a distance of 8.75 kpc and a reddening of $E(B-V) = 0.03$ mag (ref. 16), we derive $v \approx 9.6 \times 10^4 L_\odot$ pc^{-3} for the luminosity density within the density cusp (that is, for $r < 5$ arcsec; $L_\odot$, solar luminosity). Under the assumption of a mass-to-light ratio of three and a mean stellar mass of $0.5M_\odot$, this corresponds to a star number density of $n \approx 5.8 \times 10^5$ pc^{-3} (Supplementary Table 1; $M_\odot$, solar mass). Both profiles were computed with respect to the newly determined cluster centre of gravity (Supplementary Table 1), which is located ~3 arcsec southeast of the centre listed in commonly used catalogues of globular-cluster parameters[27].

Figure 2 | The BSS radial distribution. a, Cumulative radial distributions of red BSSs (red line) and blue BSSs (blue line), as functions of the projected distance, r, from the cluster centre of gravity. The distributions of subgiant-branch stars (solid black line) and horizontal-branch stars (dashed black line) are also plotted, for comparison. **b**, BSS specific frequency computed in circular areas of increasing radius r_a. The lines correspond to the overall BSS population (black) and to the red BSS (red) and blue BSS (blue) subpopulations separately. The grey area around the black line shows the 1σ uncertainty in the specific frequency. BSSs are substantially more numerous than horizontal-branch stars in the cluster centre. Although the small number of stars in the sample prevents statistical robustness in our results, we note that in the innermost 5–6 arcsec (~0.2 pc) the red BSSs tend to be as numerous as the horizontal-branch stars and dominate the ratio.

for greater radii. This suggests that dynamical friction has already affected a large portion of the cluster, with the result that almost the entire population of BSSs has sunk into the centre[8].

There is further evidence that M 30 is a highly evolved cluster from a dynamical point of view. The density profile has a steep power-law cusp in the central 5–6 arcsec (~0.2 pc; Fig. 3), suggesting that M 30 has already undergone core collapse (Supplementary Information)[9]. M 30's dynamically evolved state, combined with several suggestions[4,7,10,11] that cluster dynamics and BSS formation processes could be linked, indicates that the dual BSS sequence is probably connected to the cluster's dynamical history. In particular, during core collapse the central density rapidly increases, bringing a concomitant increase in gravitational interactions[6] able to trigger the formation of new BSSs through both direct stellar collisions and enhanced mass transfer

activity in dynamically shrunk binary systems. When considering the entire population of detected BSSs (population size, $N_{BSS} = 45$) and horizontal-branch stars ($N_{HB} = 90$), the BSS specific frequency, $F^{BSS} = N_{BSS}/N_{HB}$, is equal to 0.5, a value not particularly high in comparison with that of other clusters[12]. However, the value of F^{BSS} varies significantly over the surveyed area, reaching ~1.55 when only the central cusp of the star density profile (5–6 arcsec) is considered (Fig. 2b). So far, this is the highest value measured for the BSS specific frequency in a globular cluster, and it further supports the possibility that in M 30 we are observing the effect of enhanced gravitational interactions on single and binary stars.

To investigate this possibility, we have compared the observations with the predictions of evolutionary models of BSSs formed by direct collisions between two main-sequence stars[13] with metallicities of $Z = 10^{-4}$ and masses ranging from between $0.4M_\odot$ and $0.8M_\odot$ (thus producing BSSs with masses of between $0.8M_\odot$ and $1.6M_\odot$). As shown in Fig. 4, the blue BSS sequence is well fitted by collisional isochrones corresponding to ages of 1–2 Gyr, with the brightest blue BSS being slightly less luminous than the collision product of two turn-off-mass stars ($0.8M_\odot + 0.8M_\odot$). The observed number of stars in the blue BSS sequence is in good agreement with the expected[14]
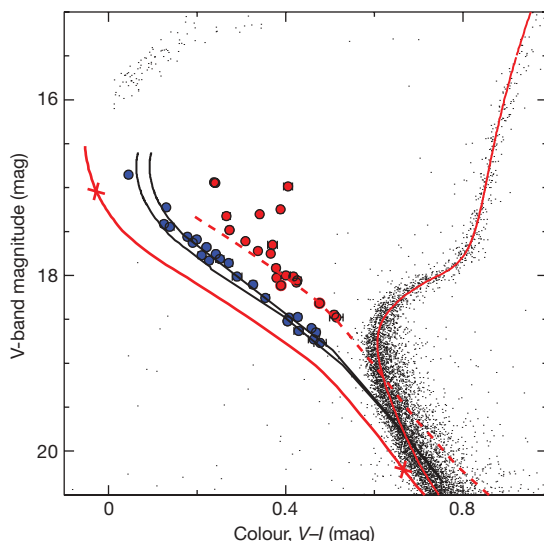
1029

Figure 4 | Comparison with collisional and binary-evolution models.
Magnified portion of the CMD of M 30. The solid black lines correspond to
the collisional isochrones[13] corresponding to 1 and 2 Gyr, respectively, which
accurately reproduce the blue BSS sequence. The solid red lines correspond
to the single-star isochrones[15] respectively corresponding to 13 Gyr (well
fitting the main cluster evolutionary sequences) and to 0.5 Gyr (representing
the reference cluster zero-age main sequence (ZAMS)). The two crosses
mark the respective positions of a $0.8M_{\odot}$ star and a $1.6M_{\odot}$ star along the
ZAMS. The dashed red line corresponds to the ZAMS shifted by 0.75 mag,
marking the position of the 'low-luminosity boundary' predicted[17] for a
population of mass-transfer binary systems. This line well reproduces the
red BSS sequence.

number of BSSs formed by direct collisions during the last 1–2 Gyr in a
cluster with total absolute magnitude comparable to that of M 30
(Supplementary Table 1). Hence, we conjecture that 1–2 Gyr ago
some dynamical process (possibly core collapse) produced the BSS
population that is now observable along the blue BSS sequence. The
origin of the red BSSs should be different, as this sequence is too red to
be properly reproduced by collisional isochrones corresponding to
any age. Figure 4 also shows the location in the CMD of single-star
isochrones[15] computed for a metallicity of $Z = 2 \times 10^{-4}$ and shifted
by the distance modulus and reddening of M 30[16]. The 13-Gyr single-
star isochrone fits the main cluster evolutionary sequences quite well,
whereas the BSS sequences are significantly offset with respect to the
0.5-Gyr single-star isochrone, which can be adopted as the cluster
ZAMS. Of particular note is the fact that the red BSS sequence is
~0.75 mag brighter than the reference ZAMS. According to the results
of recent binary-evolution models[17], during the mass-transfer phase
(which can last several gigayears, that is, a significant fraction of the
binary evolution timescale), a population of binary systems defines a
low-luminosity boundary ~0.75 mag above the ZAMS in the BSS
region (see fig. 5 of ref. 17). Hence, the BSSs that we observe along
the red BSS sequence could be binary systems still experiencing an
active phase of mass exchange.

As a result of normal stellar evolution, in a few gigayears both
collisional and mass-transfer products will populate the region
between the two sequences. The fact that we currently see two well-
separated sequences supports the hypothesis that both the blue and
the red BSS sequences were generated by a recent and short-lived
event instead of a continuous formation process. A picture in which
the two BSS sequences were generated by the same dynamical event is
therefore emerging. As suggested by the shape of the density profile
and by the location of the blue BSSs in the CMD, 1–2 Gyr ago M 30
may have undergone core collapse. This is known to increase the
gravitational interaction rate significantly and it may therefore have
boosted the formation of BSSs: the blue BSSs arise from direct stellar
collisions, whereas the red BSSs are the result of the evolution of

binary systems that first sank into the cluster centre because of
dynamical friction (or were already present in the cluster core) and
were then brought into the mass-transfer regime by hardening pro-
cesses induced by gravitational interactions during core collapse. The
detected double BSS sequence could be the signature imprinted onto
a stellar population by core collapse, with the red BSS sequence being
the outcome of the 'binary-burning' process expected to occur in the
cluster core during the late stages of the collapse[18,19]. The proposed
picture leads to a testable observational prediction: the red BSS
sequence should be populated by binaries with short orbital periods.

A recent paper[5] suggested that the dominant BSS formation channel
is the evolution of binary systems, independent of the dynamical state
of the parent cluster. Our discovery shows that binary evolution alone
does not paint a complete picture: dynamical processes can in fact play
a major part in the formation of BSSs. An appropriate survey of the
central regions of other core-collapsed clusters would help to clarify
whether the double BSS sequence is a common signature of the core-
collapse phenomenon. Moreover, detailed spectroscopic investi-
gations are worth performing to obtain a complete characterization
of the BSS properties (orbital periods, rotational velocities and so on).
In this respect, particularly promising is the search for the chemical
signature[20] of the mass-transfer process of the BSSs along the red
sequence, even though these observations represent a challenge for
the current generation of high-resolution spectrographs mounted at
8–10-m-class telescopes.

1. Marín-Franch, A. *et al.* The ACS Survey of Galactic Globular Clusters. VII. Relative
    ages. *Astrophys. J.* **694,** 1498–1516 (2009).
2. Shara, M. M., Saffer, R. A. & Livio, M. The first direct measurement of the mass of a
    blue straggler in the core of a globular cluster: BSS19 in 47 Tucanae. *Astrophys. J.*
    489, L59–L63 (1997).
3. McCrea, W. H. Extended main-sequence of some stellar clusters. *Mon. Not. R.
    Astron. Soc.* **128,** 147–155 (1964).
4. Hills, J. G. & Day, C. A. Stellar collisions in globular clusters. *Astrophys. J.* **17,** 87–93
    (1976).
5. Knigge, C., Leigh, N. & Sills, A. A binary origin for 'blue stragglers' in globular
    clusters. *Nature* **457,** 288–290 (2009).
6. Meylan, G. & Heggie, D. C. Internal dynamics of globular clusters. *Annu. Rev.
    Astron. Astrophys.* **8,** 1–143 (1997).
7. Ferraro, F. R. & Lanzoni, B. in *Dynamical Evolution of Dense Stellar Systems* (eds
    Vesperini, E., Giersz, M. & Sills, A.) 281–290 (Proc. IAU Symp. 246, Cambridge
    Univ. Press, 2008).
8. Mapelli, M. *et al.* The radial distribution of blue straggler stars and the nature of
    their progenitors. *Mon. Not. R. Astron. Soc.* **373,** 361–368 (2006).
9. Trager, S. C., Djorgovski, S. & King, I. R. in *Structure and Dynamics of Globular
    Clusters* (eds Djorgovski, S. G. & Meylan, G.) 347–355 (Astron. Soc. Pacif. Conf.
    Ser. 50, Astronomical Society of the Pacific, 1993).
10. Bailyn, C. D. Are there two kinds of blue stragglers in globular clusters? *Astrophys.
    J.* **392,** 519–521 (1992).
11. Leonard, P. J. T. Stellar collisions in globular clusters and the blue straggler
    problem. *Astron. J.* **98,** 217–226 (1989).
12. Ferraro, F. R., Sills, A., Rood, R. T., Paltrinieri, B. & Buonanno, R. Blue straggler stars:
    a direct comparison of star counts and population ratios in six Galactic globular
    clusters. *Astrophys. J.* **588,** 464–477 (2003).
13. Sills, A., Karakas, A. I. & Lattanzio, J. Blue stragglers after the main sequence.
    Astrophys. J. **692,** 1411–1420 (2009).
14. Davies, M. B., Piotto, G. & de Angeli, F. Blue straggler production in globular
    clusters. *Mon. Not. R. Astron. Soc.* **349,** 129–134 (2004).
15. Cariulo, P. Degl'Innocenti, S. & Castellani, V. Calibrated stellar models for metal-
    poor populations. *Astron. Astrophys.* **421,** 1121–1130 (2004).
16. Ferraro, F. R. *et al.* The giant, horizontal, and asymptotic branches of Galactic
    globular clusters. I. The catalog, photometric observables, and features. *Astron. J.*
    118, 1738–1758 (1999).
17. Tian, B., Deng, L., Han, Z. & Zhang, X. B. The blue stragglers formed via mass
    transfer in old open clusters. *Astron. Astrophys.* **455,** 247–254 (2006).
18. McMillian, S., Hut, P. & Makino, J. Star cluster evolution with primordial binaries. I
    – A comparative study. *Astrophys. J.* **362,** 522–537 (1990).
19. Hurley, J. *et al.* Deep Advanced Camera for Surveys imaging in the globular cluster
    NGC 6397: Dynamical models. *Astron. J.* **135,** 2129–2140 (2008).
20. Ferraro, F. R. *et al.* Discovery of carbon/oxygen-depleted blue straggler stars in 47
    Tucanae: the chemical signature of a mass transfer formation process. *Astrophys.
    J.* **647,** L53–L56 (2006).
21. Stetson, P. B. DAOPHOT – A computer program for crowded-field stellar
    photometry. *Publ. Astron. Soc. Pacif.* **99,** 191–222 (1987).

22. Stetson, P. B. The centre of the core-cusp globular cluster M15: CFHT and HST observations, ALLFRAME reductions. *Publ. Astron. Soc. Pacif.* **106**, 250–280 (1994).

23. Pietrukowicz, P. & Kaluzny, J. Variable stars in the archival HST data of globular clusters M13, M30 and NGC 6712. *Acta Astron.* **54**, 19–31 (2004).

24. Vilhu, O. Detached to contact scenario for the origin of W UMa stars. *Astron. Astrophys.* **109**, 17–22 (1982).

25. Hartigan, P. Computation of the dip statistic to test for unimodality. *Appl. Stat.* **34**, 320–325 (1985).

26. King, I. R. The structure of star clusters. III. Some simple dynamical models. *Astron. J.* **71**, 64–75 (1966).

27. Harris, W. E. A catalog of parameters for globular clusters in the Milky Way. *Astron. J.* **112**, 1487–1488 (1996).

28. Noyola, E. & Gebhart, K. Surface brightness profiles for a sample of LMC, SMC, and Fornax galaxy globular clusters. *Astron. J.* **132**, 447–466 (2006).

29. Djorgovski, S. in *Structure and Dynamics of Globular Clusters* (eds Djorgovski, S. G. & Meylan, G.) 373–382 (Astron. Soc. Pacif. Conf. Ser. 50, Astronomical Society of the Pacific, 1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions F.R.F. designed and coordinated the study. G.B., E.D. and S.B. analysed the data. A.S. and A.I.K. developed collisional models. B.L. and P.M. computed the surface density profile and performed comparisons with a single-mass King model. F.R.F. and B.L. wrote the paper. F.F.P., A.S. and R.T.R. critically contributed to discussions on, and the presentation of, the paper. All the authors contributed to discussions of the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.R.F. (francesco.ferraro3@unibo.it).

LETTERS

A binary star fraction of 76 per cent and unusual orbit parameters for the blue stragglers of NGC 188

Robert D. Mathieu[1] & Aaron M. Geller[1]

Blue straggler stars lie on or near the main sequences of star clusters (all members of which formed around the same time), but typically are more luminous than the turn-off stars and therefore long ago should have evolved off the main sequence to become giants and white dwarfs. They are thought to derive from normal main-sequence stars that have undergone a recent increase in mass. Statistical evidence indicates that in globular star clusters the blue stragglers probably form from binary stars[1]. The specific formation processes, such as mass transfer, mergers or stellar collisions during dynamical encounters of binary stars, remain unresolved. Here we report that 16 of the 21 blue stragglers (76 per cent) in the old (7-Gyr; ref. 2) open cluster NGC 188 are currently in binary systems, a frequency three times that found among normal solar-type main-sequence stars. These blue straggler binaries have a remarkable period–eccentricity distribution, with all but three having orbital periods of ~1,000 days. Moreover, these stars are rotating faster than normal main-sequence stars of the same surface temperatures. These data show that most, and possibly all, blue stragglers derive from multiple-star systems, and indicate that the several formation processes operate simultaneously. We suggest that rapid rotation of blue stragglers may place upper limits on their ages.

The colour–brightness diagram of NGC 188 shows 21 blue stragglers (Fig. 1). All are considered cluster members, on the basis of having three-dimensional velocities consistent with the distinctive bulk motion of the cluster[3,4]. As part of the WIYN Open Cluster Study[5], we have obtained highly precise ($\sigma = 0.4$ km s^{-1}) multiple radial-velocity measurements of all NGC 188 proper-motion[4] cluster members within a 30-arcmin (17-pc, 13-core-radii) radius of the cluster centre[3].

Remarkably, 16 of the 21 blue stragglers show variability in their radial velocities, indicative of them being binary stars. So far, we have obtained orbital solutions for all but one[6]. These 16 blue straggler binaries represent a short-period-binary ($P < 3,000$ d) frequency of $76 \pm 19\%$ (1σ). This frequency is not sensitive to reasonable modifications of our selection criteria for blue stragglers (Fig. 1); for example, the frequency among bright ($V < 15$ mag) blue stragglers is $76 \pm 21\%$ and the frequency among those with proper-motion membership probabilities greater than 90% is $81 \pm 23\%$. For comparison, the binary frequency among NGC 188 solar-type (14.8 mag $< V <$ 16.5 mag) main-sequence stars within the same spatial and period domain is $23 \pm 2\%$, which is broadly typical of solar-type stars[7].

Also striking are the differences between the orbital properties of the NGC 188 blue straggler and main-sequence binaries (Fig. 2). The main-sequence e–log P distribution (e, eccentricity) is typical of those found in both field and open-cluster populations of solar-type stars[7,8]. The e–log P distribution of the NGC 188 blue stragglers is strikingly different, with the orbital periods of almost all (12 of 15) of the binaries satisfying log$[P(\mathrm{d})] \approx 3$. Furthermore, among these

long-period binaries the blue straggler orbital eccentricities are smaller than among main-sequence binaries (distinct at the 98% confidence
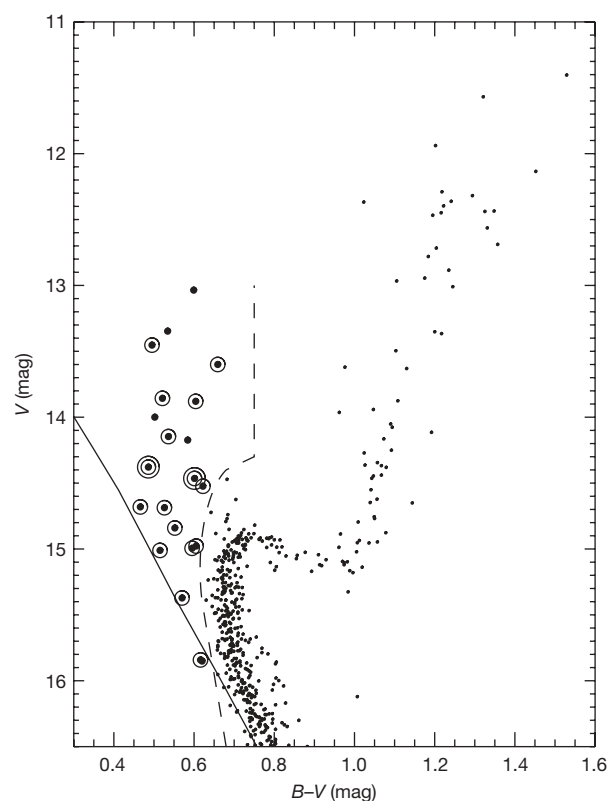

Figure 1 | Colour–brightness diagram for members of NGC 188. Brightness (increasing upwards) in V magnitudes versus colour (redder and cooler to the right) in $B-V$ magnitudes. The 21 blue stragglers are shown as larger black dots and are identified as being to the left of the dashed line at a given brightness. The line is drawn so as to not include binaries comprising two normal stars from the cluster turn-off region, which consequently are brighter in combination. For reference, we also show a zero-age main sequence[9] (solid line). Normal single stars spend most of their main-sequence lives very near this line until their core hydrogen supply is exhausted and they evolve towards the giant branch. The current turn-off mass is 1.1$M_{\odot}$ (ref. 9). Our survey sample includes 492 main-sequence stars, giants and blue stragglers. Since 1996, at least three radial-velocity measurements have been obtained for all stars, the minimum necessary to detect, through velocity variability, essentially all binary stars (>85%) with periods of <3,000 d. More than ten radial-velocity measurements are typically obtained for identified binary stars to determine the orbital parameters. The 16 blue straggler binaries are circled and the two shortest-period double-lined binaries are marked with two concentric circles, with binary 7782 the bluer of the two.

[1]Department of Astronomy, University of Wisconsin–Madison, 475 North Charter Street, Madison, Wisconsin 53706, USA.

level). The blue straggler e–$\log P$ distribution can be distinguished from the main-sequence e–$\log P$ distribution at the 99% confidence level, using a two-dimensional Kolmogorov–Smirnov test.

Each of the two short-period blue straggler binaries (binaries 5078 and 7782) is observed as a double-lined system from whose spectra we can deduce the radial velocities of both stars. (All of the other blue straggler binaries are single lined.) Analysis of binary 5078 indicates that the secondary is near the main-sequence turn-off and that the primary is a blue straggler with a dynamical mass of $1.5M_\odot$ ($M_\odot$, solar mass). The blue straggler's luminosity is less than the luminosity of a normal $1.5M_\odot$ star at any point in its evolution[9]. Binary 7782 comprises two blue stragglers of equal mass but differing surface temperatures (6,500 K and 6,325 K, respectively), indicating that mass is not the sole determinant of blue straggler properties. (See the Supplementary Information for more detailed analyses of binaries 5078 and 7782.)

These results for the NGC 188 blue stragglers are similar to those found for the blue stragglers in the old (4.5-Gyr) open cluster M 67[10]. Of 13 blue stragglers, eight (61 ± 22%) are similarly found to be short-period binaries (and the binary status of four others is unknown). Combining the two blue straggler populations, we find a short-period-binary frequency of 71 ± 14%; the excess of this binary frequency over the normal main-sequence binary frequency is valid at the 99% confidence level. The e–$\log P$ distribution of the M 67 blue stragglers also is similar to that of NGC 188, even to the extent of

having a binary with a period of 4.2 d and a modest, but non-zero, eccentricity ($e = 0.2$)[11].

Observed rotation rates are primary tests for current theories of blue straggler formation[12]. In Fig. 3, we show the distribution of $v \sin i$ (v, equatorial rotation velocity; i, inclination angle of a spin axis to our line of sight) in the colour–brightness diagram for stars in NGC 188 and in the younger (2.5-Gyr) cluster NGC 6819. We note two things: first, many of the NGC 188 blue stragglers are rotating faster than normal main-sequence stars of the same colour (surface temperature) in NGC 6819; second, there is a decrease in projected rotation velocity with decreasing surface temperature among the NCC 188 blue stragglers.

Despite many observational efforts since the identification of open-cluster blue stragglers in the mid 1950s, extensive studies have yielded very few properties that distinguish them from normal stars with similar luminosities and surface temperatures. The distributions of
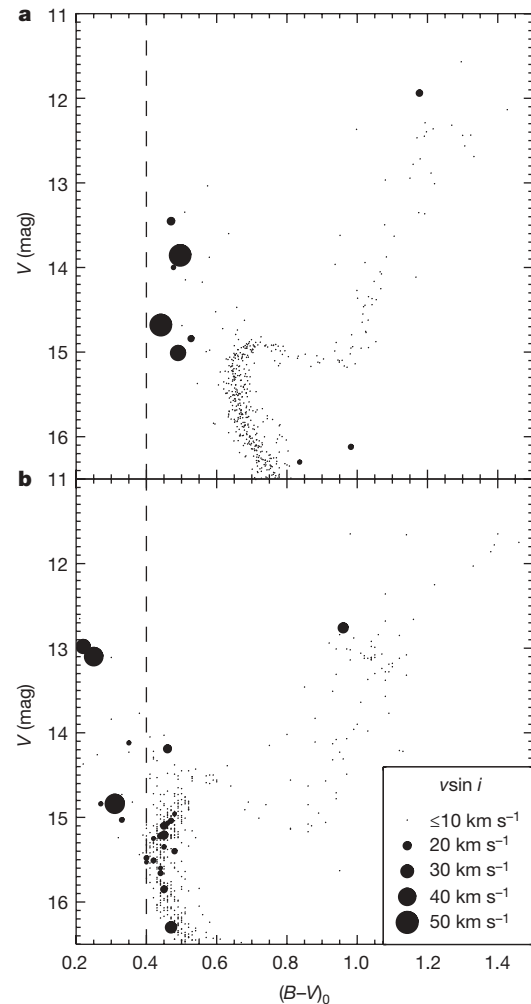

Figure 2 | NGC 188 orbital eccentricity/log period (e–$\log P$) distributions. **a**, Solar-type main-sequence binary stars. **b**, Blue straggler binary stars. Errors in eccentricity are 1σ; errors in period are smaller than the data points. The main-sequence binaries have a continuous period distribution and an eccentricity distribution that changes at the tidal-circularization cut-off period of 15 d in NGC 188[25]. Most binaries with periods less than the circularization period are tidally circularized, whereas those with longer periods have a Gaussian-like eccentricity distribution. The main-sequence e–$\log P$ distribution is very similar to that of the field binaries[7]. The blue straggler e–$\log P$ distribution is significantly different, with most of the binaries having periods of ~1,000 days and smaller eccentricities. We also note binary 5078, which has an orbital period of 4.7830 ± 0.0001 d, well below the tidal-circularization period. Its non-zero orbital eccentricity, of 0.121 ± 0.006, may be excited by a distant tertiary companion[26].

Figure 3 | Distribution of projected rotation velocity in the colour–brightness diagrams of NGC 188 and NGC 6819. **a**, NGC 188 (7 Gyr); **b**, NGC 6819 (2.5 Gyr). As in Fig. 1, we plot the V magnitude versus the $B-V$ colour, except here we have removed the interstellar reddening from the colour for each cluster and thus plot the intrinsic $(B-V)_0$ colour. We plot only cluster members in both panels. The point sizes correspond to the projected rotational velocities ($v \sin i$) for each star, as indicated. Given our spectral resolution of 15 km s^{-1}, we are able to measure $v \sin i$ values in excess of 10 km s^{-1} with a typical precision of 1.5 km s^{-1}. For each cluster, we have removed close binaries whose periods are less than the corresponding tidal-circularization periods, so the $v \sin i$ values shown have not been affected by tidal evolution. The dashed line at $(B-V)_0 = 0.4$ is to guide the eye in comparing the $v \sin i$ values of stars of similar colours in the two clusters.

blue stragglers in colour–brightness diagrams have been the primary interface with theories of their formation and evolution. Thus, the binary and rotation properties of the blue stragglers in NGC 188 reported here are among the most distinguishing properties of open-cluster blue stragglers.

The very high binary frequency points definitively to an origin in multiple-star systems for most of the NGC 188 blue stragglers. Formation processes involving only single stars, such as unusual internal mixing or direct stellar collisions, would not in themselves produce blue stragglers in binary systems. Furthermore, cluster simulations show that it is very unlikely that 76% of a population of single blue stragglers could be dynamically exchanged into pre-existing binary stars[13]. Finally, our observation techniques can miss companions with periods of a few thousand days or longer; it thus remains possible that all of the NGC 188 blue stragglers are in binaries and formed in multiple-star systems.

Currently discussed formation channels for blue straggler binaries include mass transfer, mergers in hierarchical multiple systems and direct stellar collisions occurring during dynamical encounters of binary stars (or higher-order multiple systems). The presence of blue stragglers in NGC 188 and M 67 suggest that all of these processes are operating.

Three of the 12 blue straggler binaries with periods of ~1,000 d in NGC 188 have orbital eccentricities indistinguishable from zero (circular orbits), but no circular orbits are found among the main-sequence binaries with similar periods. Circular orbits at such periods are indicative of tidal circularization involving giant stars[14,15]. Thus, the circular orbits of these blue stragglers point towards mass transfer from a prior giant star enhancing the mass of the current blue straggler. A prediction of this interpretation is that the secondary stars of these binaries are stellar remnants.

Recently a new channel for blue straggler formation was suggested, in which the Kozai cycle induces a common envelope or a merger of the inner binary in a hierarchical triple system[16]. (See the Supplementary Information for further information.) The product of this process would be a blue straggler in a wide binary, with the original tertiary star as the companion. Thus, this new channel for merger formation of blue stragglers produces a high binary frequency and possibly an e–log P distribution similar to that found here. Alternatively, we note that triple systems have larger cross-sections for dynamical encounters than do the close binaries within them; we conjecture that resonant dynamical encounters may similarly tighten the close binaries and produce blue stragglers.

The blue straggler S1082 in M 67 is perhaps the most compelling candidate for a blue straggler derived from stellar collisions within a binary–binary encounter[17,18]. This very probably triple system consists of two blue stragglers, one of which is in a very close (1.07-d period) eclipsing binary. The eclipsing-binary blue straggler has a directly measured mass of $2.5M_\odot$ and the secondary star has a mass of $1.6M_\odot$. As the turn-off mass in M 67 is $1.3M_\odot$, mass from at least four stars is needed to produce this binary. Furthermore, the secondary star currently lies within the main sequence despite its high mass, indicating that it is not in equilibrium and was thus recently formed. Finally, the short-period orbit suggests that the binary was dynamically formed. All of these properties point to collision origins through dynamical encounters of multiple-star systems. We expect that the NGC 188 blue straggler binary 7782 was also formed through dynamical encounters, although the blue stragglers may have formed independently and come together through a dynamical exchange.

An important population for comparison are the field blue stragglers, which are probably the result of mass transfer rather than stellar dynamical encounters. Their e–log P distribution is rather similar to those found in NGC 188 and M 67[19]. We suggest that a critical characteristic of all blue straggler binary populations is the frequency of main-sequence secondary stars. The number of single-lined blue straggler binaries in NGC 188 and M 67 marginally rules out the hypothesis

that all have main-sequence secondary stars. Spectra for each with high signal-to-noise ratios will give definitive evidence.

Within a star cluster, these several formation processes, their rates and their products are further modified by dynamical interactions. Recent numerical N-body simulations of open-cluster evolution include both gravitational dynamics and stellar evolution processes, and thereby produce blue stragglers[20]. The most comprehensive cluster simulation so far specifically sought to produce the old open cluster M 67[21]. Here we use this simulation as a proxy for NGC 188 as well.

Given the age of M 67, the number of blue stragglers (20) in this simulation is in reasonable agreement with the observed numbers. However, the frequency of blue straggler binaries is a factor of two to three less than that observed in M 67 or NGC 188 (Supplementary Information). This lower binary frequency may result from the lack of primordial triples in the simulation.

Furthermore, the simulated blue straggler binaries have a broader period distribution and tend towards higher eccentricities than are observed in either cluster. Both the high orbital eccentricities and the broad period distribution were established by dynamical interactions with other cluster stars, suggesting that the simulations are more dynamically active than the actual clusters. Indeed, the observation of binaries with circular orbits and periods of ~1,000 d in both NGC 188 and M 67 indicates that these blue straggler binaries have not been disturbed dynamically since their formation.

We find that many NGC 188 blue stragglers are rotating more rapidly than normal main-sequence stars of the same surface temperatures. Nonetheless, they are not rotating as fast as predicted by collision and merger models of blue stragglers[12]. We also find that the blue straggler rotation rates decrease with decreasing surface temperature. A similar trend exists among normal main-sequence stars spanning the same temperature range[22]. The spin-down of cooler stars is associated with the formation of surface convective zones and magnetic fields, and the onset of stellar winds with decreasing surface temperature; the magnetized winds transport away angular momentum on timescales of several hundred megayears[23,24].

Whether blue stragglers of these same surface temperatures also have surface convection zones or magnetic fields is unknown; some proposed formation mechanisms may exclude their presence[12]. If the blue stragglers do not have effective mechanisms by which to lose angular momentum, then serious challenges are placed before formation mechanisms—such as collisions and mergers—that are thought to produce very rapidly rotating products from the pre-event orbital angular momentum. Alternatively, if the internal structure, magnetic fields and winds of blue stragglers mimic normal main-sequence stars, with commensurate spin-down times, then our finding that some blue stragglers have not yet spun down will place upper limits on their ages.

1. Knigge, A., Leigh, N. & Sills, A. A binary origin for 'blue stragglers' in globular clusters. *Nature* **457**, 288–290 (2009).
2. Sarajedini, A., von Hippel, T., Kozhurina-Platais, V. & Demarque, P. WIYN Open Cluster Study. II. UBVRI CCD photometry of the open cluster NGC 188. *Astron. J.* **118**, 2894–2907 (1999).
3. Geller, A. M., Mathieu, R. D., Harris, H. C. & McClure, R. D. WIYN Open Cluster Study. XXXII. Stellar radial velocities in the old open cluster NGC 188. *Astron. J.* **135**, 2264–2278 (2008).
4. Platais, I., Kozhurina-Platais, V., Mathieu, R. D., Girard, T. M. & van Altena, W. F. WIYN Open Cluster Study. XVII. Astrometry and membership to V = 21 in NGC 188. *Astron. J.* **126**, 2922–2935 (2003).
5. Mathieu, R. D. in *Stellar Clusters and Associations: Convection, Rotation and Dynamos* (eds Pallavicini, R., Micela, G. & Sciortino, S.) 517–531 (Astron. Soc. Pacif. Conf. Ser. 198, Astronomical Society of the Pacific, 2000).
6. Geller, A. M., Mathieu, R. D., Harris, H. C. & McClure, R. D. WIYN Open Cluster Study. XXXVI. Spectroscopic binary orbits in NGC 188. *Astron. J.* **137**, 3743–3760 (2009).
7. Duquennoy, A. & Mayor, M. Multiplicity among solar-type stars in the solar neighbourhood. II. Distribution of the orbital elements in an unbiased sample. *Astron. Astrophys.* **248**, 485–524 (1991).
8. Meibom, S. & Mathieu, R. D. A robust measure of tidal circularization in coeval binary populations: the solar-type spectroscopic binary population in the open cluster M35. *Astrophys. J.* **620**, 970–983 (2005).

9. Girardi, L. *et al.* Theoretical isochrones in several photometric systems. I. Johnson-Cousins-Glass, HST/WFPC2, HST/NICMOS, Washington, and ESO Imaging Survey filter sets. *Astron. Astrophys.* **391**, 195–212 (2002).

10. Latham, D. W. Spectroscopic binaries in M67. *Highlights Astron.* **14**, 444–445 (2007).

11. Milone, A. A. E. & Latham, D. W. in *Evolutionary Processes in Interaction Binary Stars* (eds Kondo, Y., Sisteró, R. F. & Polidan, R. S.) 475–478 (Proc. IAU Symp. 151, Springer, 1992).

12. Sills, A., Adams, T. & Davies, M. B. Blue stragglers as stellar collision products: the angular momentum question. *Mon. Not. R. Astron. Soc.* **358**, 716–725 (2005).

13. Portegies Zwart, S., Hut, P., McMillan, S. L. W. & Verbunt, F. Star cluster ecology II. Binary evolution with single-star encounters. *Astron. Astrophys.* **328**, 143–157 (1997).

14. Mermilliod, J.-C. & Mayor, M. in *Binaries as Tracers of Stellar Formation* (eds Duquennoy, A. & Mayor, M.) 183–201 (Cambridge Univ. Press, 1992).

15. Mathieu, R. D., Latham, D. W. & Griffin, D. W. Orbits of 22 spectroscopic binaries in the open cluster M67. *Astron. J.* **100**, 1859–1881 (1990).

16. Perets, H. B& Fabrycky, D. C. On the triple origin of blue stragglers. *Astrophys. J.* **697**, 1048–1056 (2009).

17. Sandquist, E. L., Latham, D. W., Shetrone, M. D. & Milone, A. A. E. The blue straggler RS Canum Venaticorum star S1082 in M67: a detailed light curve and the possibility of a triple. *Astron. J.* **125**, 810–824 (2003).

18. van den Berg, M., Orosz, J., Verbunt, F. & Stassun, K. The blue straggler S1082: a triple system in the old open cluster M 67. *Astron. Astrophys.* **375**, 375–386 (2001).

19. Carney, B. W., Latham, D. W., Laird, J. B., Grant, C. E. & Morse, J. A. A survey of proper-motion stars. XIV. Spectroscopic binaries among metal-poor field blue stragglers. *Astron. J.* **122**, 3419–3435 (2001).

20. Hurley, J. R., Tout, C. A. & Pols, O. R. Evolution of binary stars and the effect of tides on binary populations. *Mon. Not. R. Astron. Soc.* **329**, 897–928 (2002).

21. Hurley, J. R., Pols, O. R., Aarseth, S. J. & Tout, C. A. A complete N-body model of the old open cluster M67. *Mon. Not. R. Astron. Soc.* **363**, 293–314 (2005).

22. Kraft, R. P. Studies of stellar rotation. V. The dependence of rotation on age among solar-type stars. *Astrophys. J.* **150**, 551–570 (1967).

23. Skumanich, A. Time scales for Ca II emission decay, rotational braking, and lithium depletion. *Astrophys. J.* **171**, 565–567 (1972).

24. Meibom, S. in *The Ages of Stars* (eds Mamajek, E. E., Soderblom, D. R. & Wyse, R. F. G.) 357–362 (Proc. IAU Symp. 258, Cambridge Univ. Press, 2009).

25. Mathieu, R. D., Meibom, S. & Dolan, C. J. WIYN Open Cluster Study. XVIII. The tidal circularization cutoff period of NGC 188. *Astrophys. J.* **602**, L121–L123 (2004).

26. Mazeh, T. Eccentric orbits in samples of circularized binary systems: the fingerprint of a third star. *Astron. J.* **99**, 675–677 (1990).

Author Contributions R.D.M. and A.M.G. contributed equally to this work.

LETTERS

A strong, highly-tilted interstellar magnetic field near the Solar System

M. Opher[1], F. Alouani Bibi[1], G. Toth[2], J. D. Richardson[3], V. V. Izmodenov[4] & T. I. Gombosi[2]

Magnetic fields play an important (sometimes dominant) role in the evolution of gas clouds in the Galaxy, but the strength and orientation of the field in the interstellar medium near the heliosphere has been poorly constrained. Previous estimates of the field strength range from 1.8–2.5 μG and the field was thought to be parallel to the Galactic plane[1] or inclined by 38–60° (ref. 2) or 60–90° (ref. 3) to this plane. These estimates relied either on indirect observational inferences or modelling in which the interstellar neutral hydrogen was not taken into account. Here we report measurements of the deflection of the solar wind plasma flows in the heliosheath[4] to determine the magnetic field strength and orientation in the interstellar medium. We find that the field strength in the local interstellar medium is 3.7–5.5 μG. The field is tilted ~20–30° from the interstellar medium flow direction (resulting from the peculiar motion of the Sun in the Galaxy) and is at an angle of about 30° from the Galactic plane. We conclude that the interstellar medium field is turbulent or has a distortion in the solar vicinity.

The local interstellar medium magnetic field (B_{ISM}) is one of the key elements that control the interaction between the Solar System and the interstellar medium. Determining its strength and orientation is crucial because B_{ISM} affects the shape of the Solar System and the filtration of particles that stream into the Solar System from the interstellar medium. However, until the Voyager spacecraft crosses the heliopause, we cannot directly measure the orientation or strength of B_{ISM}.

Previous work estimated the orientation and intensity of B_{ISM}, but with large uncertainties. Measurements of the polarization of light from nearby stars[1] suggest that the average field over spatial scales of parsecs is parallel to the Galactic disk, but gives no information on the local field direction. The backscattered solar Lyman-α radiation[2] gave a field direction inclined 38–60° with respect to the Galactic plane with the angle between the velocity of the interstellar medium and magnetic field assumed to be 30–60°. This method relies on the assumption that B_{ISM} is in a plane defined by the interstellar H and He flow directions as they penetrate the Solar System. However, recent studies have shown that this assumption may not be valid[5].
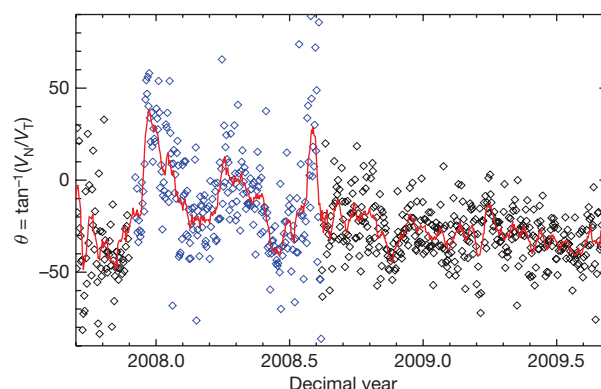

Figure 1 | Flows in the heliosheath as measured by Voyager 2. The flow angle $\theta = \tan^{-1}(V_N/V_T)$ is shown from day 277 of 2007 to day 245 of 2009 as measured by Voyager 2 in the R–T–N coordinate system. The horizontal axis shows the time in decimal years (day/365) from 2007–2009. The period between 2007.95 and 2008.62 (blue points) seems to be dominated by transients that may be driven by interactions between the fast and slow solar wind. We model the steady-state heliosphere and thus remove this region to calculate the average flow angle in the heliosheath. The average angle θ is $-29° \pm 1.0°$. The red line shows 11-day running averages of the angles.

The deflection of the H from the He flow direction is affected both by the orientation and by the strength of B_{ISM}. The same observed average deflection can be produced by different orientations and intensities of B_{ISM}. Previous studies do not constrain the strength of the field. We previously[3,6,7] derived the strength and orientation of the B_{ISM} from a numerical model constrained by Voyager 1 and 2 particle streaming and radio observations . This model did not include neutral H atoms, so it underestimated the strength of the interstellar magnetic field and had large uncertainties in the field direction[8,9].

Here we constrain the orientation and strength of B_{ISM} using a new data set, the heliosheath flows obtained from Voyager 2 measurements.

Table 1 | Definitions of the parameters used

α	Angle between the local interstellar magnetic field and interstellar wind (see also Fig. 3 inset)
β	Angle between the solar equatorial plane and the plane defined by the interstellar magnetic field and the interstellar wind
θ	$\tan^{-1}(V_N/V_T)$; the angle describing the flow deflection in the heliosheath
θ_1	Angle between the shock normal n and the Sun-centred radius vector
ϕ_1	Angle measured between the tangential direction T and the projection of n in the N–T plane (see Supplementary Information for a diagram)
V_S	Shock velocity
R_{TS}	Distance of the termination shock to the Sun
V_R	Radial velocity in the R–T–N coordinate system, that is, a local cartesian system centred at the spacecraft. R is radially outward from the Sun, T is in the plane of the solar equator and positive in the direction of solar rotation, and N completes a right-handed system
V_N	Normal velocity in the R–T–N coordinate system
V_T	Tangential velocity in the R–T–N coordinate system

[1]George Mason University, 4400 University Drive, Fairfax, Virginia 22030, USA. [2]Center for Space Environment Modeling, University of Michigan, Ann Arbor, Michigan 48109, USA. [3]Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 37-655, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [4]Lomonosov Moscow State University, Space Research Institute (IKI) and Institute for Problems in Mechanics, Russian Academy of Science, 84/32 Profsoyuznaya Street, Moscow 117997, Russia.
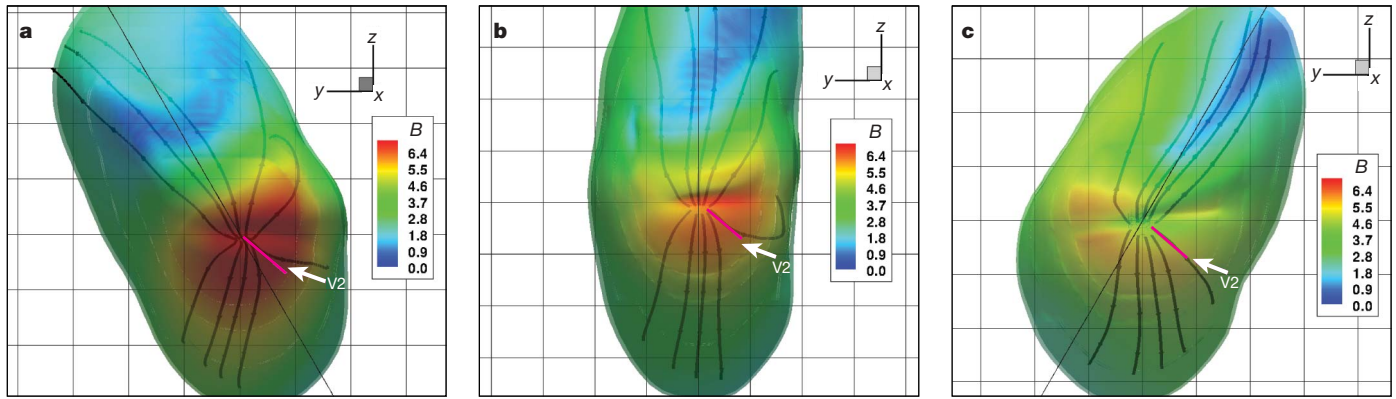
Figure 2 | Flows in the heliosheath for different B_{ISM} directions. The shapes of the heliopause are shown for $B_{ISM} = 4.4\,\mu G$, $\alpha = 45°$ and β equal to $60°$ (**a**), $90°$ (**b**) and $120°$ (**c**). The colours indicate the total magnetic field strength (in microgauss) and are shown on the iso-surface of the heliopause defined for $\ln T = 11.5$, where T is the temperature. The viewpoint is from the $+x$ direction. The streamlines show qualitatively the difference in flow patterns in the heliosheath. The interstellar wind is along the x-axis, the solar rotation axis is along the z-axis and the y-axis completes the right-handed coordinate system. In this coordinate system, the Voyager 2 trajectory is shown by the

pink lines labelled V2. The Voyager 2 ecliptic latitude and longitude are $-31.2°$ and $213.4°$ respectively, neglecting the $7.25°$ tilt of the solar equator with respect to the ecliptic plane. The heliopause surfaces and flows for the $\beta = 60°$ (**a**) and $\beta = 120°$ (**b**) orientations are almost mirror images. The heliopause is asymmetric, both in the north–south and in the east–west directions, and has a plane of symmetry approximately parallel to the plane of the local B_{ISM}. The intensity of the magnetic field outside the heliopause is larger in the southern solar hemisphere rather than in the northern solar hemisphere (**a**) owing to the slowing-down of the plasma flow.

The heliosheath flow angle $\theta = \tan^{-1}(V_N/V_T)$ depends on the interstellar medium magnetic field. (The coordinate system is defined in Table 1.) Figure 1 shows the flow angle θ in the heliosheath at Voyager 2. This dependence of the flow angle on the magnetic field orientation can be seen qualitatively in Fig. 2, which shows the flows predicted by a three-dimensional magnetohydrodynamic model with five fluids (one ionized and four neutral H fluids; see description in Supplementary Information).

The orientation of the interstellar magnetic field is defined by the angles α and β, where α is the angle between the interstellar magnetic field and the interstellar wind and β is the angle between the solar equator and the plane defined by the interstellar magnetic field and the interstellar wind (see inset in Fig. 3). (The Local Standard of Rest

Figure 3 | Modelled flows as a function of the B_{ISM}. The flow angle $\theta = \tan^{-1}(V_N/V_T)$ is shown as a function of the B_{ISM} orientation. Flow angles for $\alpha = 20°$, $30°$, $45°$ and $60°$ are shown in blue, purple, brown and green, respectively. The size of the squares corresponds to the intensity of B_{ISM} from 2.5–$5.5\,\mu G$. The model flow angle θ was estimated using all the points within two cells of resolution (or 6 AU) of the Voyager 2 trajectory (see Supplementary Information). The error bars are the root mean square of all the values. The observed flow angle $\theta = -29° \pm 1.0°$ is shown by the dotted line. The blue dashed line illustrates the dependence of the flow angle on β for $\alpha = 20$–$30°$. The inset shows the orientation of α and β in the model coordinate systems (for more details see Table 1). The velocity of the interstellar wind is indicated by the label V_{ISM}.

is the mean motion of the stars and gas in the solar neighbourhood in rotation around the Galactic centre. The Sun moves relative to the Local Standard of Rest at $13.4\,km\,s^{-1}$ in the direction of Galactic rotation, $-9\,km\,s^{-1}$ in the radial direction, and $3.7\,km\,s^{-1}$ in the vertical direction. The cloud of material that surrounds the Sun moves in the Local Standard of Rest in a nearly perpendicular motion. The result of these two motions is that we observe interstellar material flowing towards the Sun, called the interstellar wind. We use the interstellar wind direction to define the x-axis of our model (see Supplementary Fig. 4.) We constrain the value of α by comparing the observed distances of the termination shock from the Sun in the Voyager 1 and Voyager 2 directions[10,11] to the model results (Table 2) and find that values of α between $20°$ and $30°$ fit the data. The magnitude of B_{ISM} must be between $3.7\,\mu G$ and $5.5\,\mu G$ to match the observed termination shock crossing distances. We note that other work found α to be 15–$20°$ and the magnitude to be about $4\,\mu G$, but these studies did not take into account the interplanetary magnetic field[12]. We constrain the value of β by comparing the measured and simulated flow velocities. Figure 3 shows the dependence of β on the flow angle for several values of α. For $\alpha = 20$–$30°$, as β increases the predicted flow angle θ agrees better with the measured angle. β has to be less than $90°$ due to magnetic connectivity[3], so β near $90°$ (or about $30°$ from the Galactic plane) gives the best fit to the observations.

We analyse the effects of non-stationarity and numerical resolution on our estimates of the magnetic field orientation and find them to be negligible. The effects of the non-stationarity of the shock can be obtained from the theoretical predictions of θ for a three-dimensional magnetized moving shock given by the Rankine–Hugoniot conditions and described by the angles θ_1 and ϕ_1 (for definitions see Table 1). The termination shock at the time of the Voyager 2 crossing had, on average, $\theta_1 = 15°$, $\phi_1 = 165°$ and $V_S = 88.09\,km\,s^{-1}$, where V_S is the shock velocity (see Supplementary Information for more details). Using the Rankine–Hugoniot conditions with these values for θ_1 and ϕ_1, we estimate that the dependence of θ (the orientation of the flow which constrains the orientation of the magnetic field) on the shock velocity V_S is less than $0.1°$. In the Supplementary Information we show that the effect of numerical resolution is also small (less than $8°$). We show that the obliquity of the shock given by θ_1 and ϕ_1 depends on the interstellar magnetic field (see Supplementary Information for details).

These results show that the local B_{ISM} orientation with respect to the Galactic plane is different from that determined from ground-based
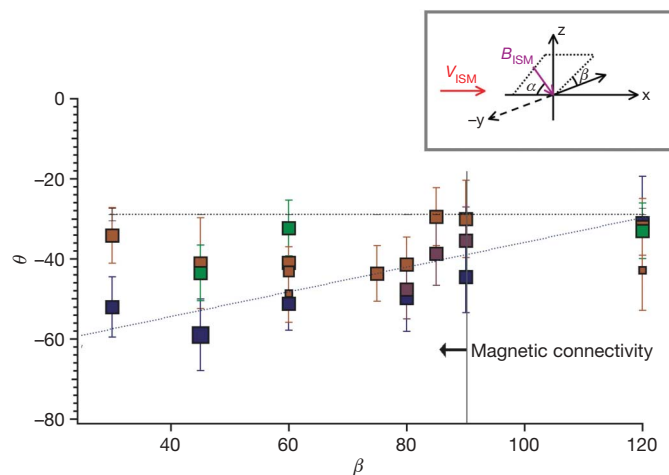
Table 2 | Termination shock distances at Voyager 1 and 2 as a function of B_{ISM}

$B\,(\mu G)$	$\beta\,(°)$	$\alpha\,(°)$	$R_{TS}(V1)\,(AU)$	$R_{TS}(V2)\,(AU)$	$R_{TS}(V2) - R_{TS}(V1)\,(AU)$
4.4	30	20	88.7 ± 3.6	79.8 ± 2.3	8.9
4.4	60	20	90.3 ± 4.0	81.2 ± 2.7	9.1
4.4	80	20	94.1 ± 4.8	85.8 ± 2.7	8.3
4.4	120	20	94.2 ± 3.4	92.9 ± 4.7	1.3
5.5	45	20	84.6 ± 3.6	72.4 ± 2.6	12.2
4.4	80	30	88.5 ± 3.8	79.3 ± 2.4	9.2
4.4	85	30	88.8 ± 3.8	82.6 ± 4.9	6.2
4.4	90	30	89.2 ± 4.2	81.1 ± 2.5	8.1
2.5	60	45	89.1 ± 3.8	85.3 ± 3.0	3.8
3.7	60	45	82.7 ± 3.5	70.7 ± 3.4	12.0
4.4	60	45	76.6 ± 1.5	67.8 ± 2.3	8.8
4.4	75	45	78.1 ± 2.0	70.2 ± 2.4	7.9
4.4	80	45	79.2 ± 2.6	71.8 ± 1.9	7.4
4.4	85	45	80.2 ± 1.9	74.5 ± 2.6	7.8
4.4	120	45	80.3 ± 3.0	78.9 ± 2.6	1.9
4.4	30	45	73.1 ± 1.9	67.1 ± 2.0	6.0
4.4	45	45	74.4 ± 2.0	68.4 ± 2.6	6.0
4.4	90	45	80.4 ± 1.9	73.3 ± 2.0	7.1
4.4	120	60	72.9 ± 1.9	67.2 ± 2.2	5.7
4.4	60	60	73.0 ± 2.0	67.5 ± 2.2	5.5

The error bars in the termination shock distances R_{TS} are the average of the upstream and downstream distances. The angle α is the dominant factor in determining the distance of the termination shock to the Sun and the asymmetry between Voyager 1 (V1) and Voyager 2 (V2), $R_{TS}(V1) - R_{TS}(V2)$ (in astronomical units, AU). Models that take into account the motion of the shock[22,23] estimate an asymmetry of 7 AU between the termination shock distances in the Voyager 1 and 2 directions. To reproduce the observed asymmetry of 7 AU, B_{ISM} must be >3.7 μG. To reproduce the distances $R_{TS}(V1) = 91$ AU and $R_{TS}(V2) = 83.7$ AU, the angle α has to be 20–30°.

measurements, which determine the average field orientation over large distances[1]. This difference could be a result of turbulence in the interstellar medium[13,14] which could cause the local magnetic field direction to differ dramatically from that of the large-scale field. This difference could also be a consequence of local distortion of the magnetic field in the solar vicinity[15]. Within 100–200 pc from the Sun, the interstellar gas is embedded in the Local Bubble[16], a huge region of hot tenuous plasma which contains small, cooler, denser clouds such as the Local Interstellar Cloud (also called the Local Cloud), which envelops the Sun. Collisions between the Local Cloud and other cooler, denser clouds in the Local Bubble, such as the G cloud[17], could cause such a distortion. Theories for the creation of the Local Cloud suggest that the magnetic fields are strong, of the order of 4–7 μG (ref. 18).

For a magnetic field strength in the Local Cloud of 4 μG, which we report here, the magnetic pressure would be $P_B/k \approx 4{,}500\,\mathrm{K\,cm}^{-3}$ (where k is the Boltzmann constant), almost twice the Local Cloud thermal pressure of $P_{th}/k \approx 2{,}500\,\mathrm{K\,cm}^{-3}$ (refs 19, 20). If this magnetic field strength were representative of the local field over a scale of about 10 pc, the magnetic energy density could dominate the energy budget in the Local Cloud[20,21].

1. Frisch, P. C. LISM structure-fragmented superbubble shell? *Space Sci. Rev.* **78**, 213–222 (1996).
2. Lallement, R. *et al.* Deflection of the interstellar neutral hydrogen flow across the heliospheric interface. *Science* **307**, 1447–1449 (2005).
3. Opher, M., Stone, E. C. & Gombosi, T. I. The orientation of the local interstellar magnetic field. *Science* **316**, 875–878 (2007).
4. Richardson, J. D. *et al.* Cool heliosheath plasma and deceleration of the upstream solar wind at the termination shock. *Nature* **454**, 63–66 (2008).
5. Pogorelov, N. V. & Zank, G. P. The direction of the neutral hydrogen velocity in the inner heliosphere as a possible interstellar magnetic field compass. *Astrophys. J.* **636**, L161–L164 (2006).
6. Opher, M., Stone, E. C. & Liewer, P. C. The effects of a local interstellar magnetic field on Voyager 1 and 2 observations. *Astrophys. J.* **640**, L71–L74 (2006).
7. Opher, M., Richardson, J. C., Toth, G. & Gombosi, T. I. Confronting observations and modeling: the role of the interstellar magnetic field in Voyager 1 and 2 asymmetries. *Space Sci. Rev.* **143**, 43–55 (2009).
8. Pogorelov, N. V., Stone, E. C., Florinski, V. & Zank, G. P. Termination shock asymmetries as seen by the Voyager spacecraft: the role of the interstellar magnetic field and neutral hydrogen. *Astrophys. J.* **668**, 611–624 (2007).
9. Pogorelov, N. V., Heerikhuisen, J. & Zank, G. P. Probing heliospheric asymmetries with an MHD-kinetic model. *Astrophys. J.* **675**, L41–L44 (2008).
10. Stone, E. C. *et al.* An asymmetric solar wind termination shock. *Nature* **454**, 71–74 (2008).
11. Stone, E. C. *et al.* Voyager 2 explores the termination shock region and the heliosheath beyond. *Science* **309**, 2017–2020 (2005).
12. Izmodenov, V. V. Local interstellar parameters as they are inferred from analysis of observations inside the heliosphere. *Space Sci. Rev.* **143**, 139–150 (2009).
13. Minter, A. H. & Spangler, S. R. Observation of turbulent fluctuations in the interstellar plasma density and magnetic field on spatial scales of 0.01 to 100 parsecs. *Astrophys. J.* **458**, 194–214 (1996).
14. Jokipii, J. R. Our interstellar neighborhood. *Science* **307**, 1424–1425 (2007).
15. Frisch, P. C. *et al.* The galactic environment of the Sun: interstellar material inside and outside the heliosphere. *Space Sci. Rev.* **146**, 235–273 (2009).
16. Linsky, J. L. Solving the mysteries of the diffuse interstellar medium with high-resolution UV spectroscopy. *Astrophys. Space Sci.* **320**, 85–90 (2009).
17. Redfield, S. & Linsky, J. The structure of the local Interstellar Medium. IV. Dynamics, morphology, physical properties, and implications of cloud-cloud interactions. *Astrophys. J.* **673**, 283–314 (2008).
18. Cox, D. & Helenius, L. Flux-tube dynamics and a model for the origin of the local fluff. *Astrophys. J.* **583**, 205–228 (2003).
19. Jenkins, E. B. & Tripp, T. M. The distribution of thermal pressures in the interstellar medium from a survey of CI fine-structure excitation. *Astrophys. J. Suppl. Ser.* **137**, 297–340 (2001).
20. Shelton, R. *et al.* The local bubble debate. Report from sessions 1 and 3. *Space Sci. Rev.* **143**, 303–309 (2009).
21. Jenkins, E. B. Pressure and ionization balances in the circum-heliospheric interstellar medium and the local bubble. *Space Sci. Rev.* **143**, 205–216 (2009).
22. Izmodenov, V. V., Malama, Y. G. & Ruderman, M. S. Modeling of the outer heliosphere with the realistic solar cycle. *Adv. Space Res.* **41**, 318–324 (2008).
23. Richardson, J. D., Liu, Y. & Wang, C. Solar wind structure in the outer heliosphere. *Adv. Space Res.* **41**, 237–244 (2008).

Observation of molecular orbital gating

Hyunwook Song[1,2], Youngsang Kim[3]†, Yun Hee Jang[2], Heejun Jeong[3], Mark A. Reed[4] & Takhee Lee[1,2]

The control of charge transport in an active electronic device depends intimately on the modulation of the internal charge density by an external node[1]. For example, a field-effect transistor relies on the gated electrostatic modulation of the channel charge produced by changing the relative position of the conduction and valence bands with respect to the electrodes. In molecular-scale devices[2–10], a longstanding challenge has been to create a true three-terminal device that operates in this manner (that is, by modifying orbital energy). Here we report the observation of such a solid-state molecular device, in which transport current is directly modulated by an external gate voltage. Resonance-enhanced coupling to the nearest molecular orbital is revealed by electron tunnelling spectroscopy, demonstrating direct molecular orbital gating in an electronic device. Our findings demonstrate that true molecular transistors can be created, and so enhance the prospects for molecularly engineered electronic devices.

The demonstration of a true three-terminal molecular device, one that depends on external modulation of molecular orbitals, has been the outstanding challenge of the field of molecular electronics since soon after its inception[5], but until now it has defied experimental efforts. Theoretical proposals indicate that field-effect gating of a molecular junction is possible[6,7]. Carrier modulation in molecular junctions can be achieved by alternative non-molecular mesoscopic mechanisms such as Coulomb blockade and Kondo effects[8–10], and has been claimed to be observed in an electrochemical break junction[11], but until now proof of direct orbital gate control of a solid-state device has been elusive. Single-molecule orbital modulation has been demonstrated in two-terminal, non-device scanning tunnelling microscopy experiments[12], lending encouragement to the development of three-terminal device implementations, notwithstanding concerns of the degree of gate control that can be achieved in a field-effect transistor (FET) configuration[6,13]. Here we demonstrate direct electrostatic modulation of orbitals in a solid-state molecular FET configuration, with both effective gate control and enhanced resonant coupling of the molecular orbitals to the source and drain electrodes. We show this using a multiprobe approach combining a variety of transport techniques that give a self-consistent characterization of the molecular junction.

As illustrated in the inset of Fig. 1a, individual molecules are connected to source and drain electrodes with a bottom-gate control electrode in a FET configuration. In such devices, the energies of the molecular orbitals with respect to the Fermi level of the electrodes can be directly tuned by adjusting a gate voltage, V_G. We make such devices using the well-known electromigration technique of fracturing a continuous gold wire (coated with the desired molecules, in vacuum, at 4.2 K) that is placed over an oxidized aluminium gate electrode[8,9] (Methods Summary). This produces source and drain electrodes with a nanometre-scale gap, which are often bridged by single or very few molecules, creating molecular junctions. We have examined a sample size of 418 devices containing two prototype molecules: the control, a 1,8-octanedithiol (ODT) with an alkyl

backbone as a σ-saturated aliphatic molecule; and the active device, a 1,4-benzenedithiol (BDT) with a delocalized aromatic ring as a π-conjugated molecule. We found 35 devices with sigmoidal and significantly gate-dependent source–drain current (I)–voltage (V) characteristics. The measured low-bias conductance is in good agreement with single-molecule conductance values of the molecules of interest (Supplementary Fig. 4). Coherent, off-resonance tunnelling with a strong molecule–metal contact coupling (for example, the thiol–gold bond) can be reasonably expected as a conduction mechanism for these junctions, in accordance with the form of the symmetric $I(V)$ curves[14,15], the lack of a significant conductance gap[14,15], the exponential dependence of conductance on molecular length[15–17] (Supplementary Fig. 6) and the temperature-independent $I(V)$ behaviour[17] (Supplementary Fig. 5).

Figure 1a shows representative $I(V)$ curves for a Au–ODT–Au (σ-saturated aliphatic) junction measured at different V_G values. The tunnelling current passing through the ODT junction increases as V_G becomes increasingly negative. The corresponding analysis of $\ln(I/V^2)$ versus $1/V$, the graph of which is known as a Fowler–Nordheim plot, is shown in Fig. 1b. Here two distinct transport regimes are evident; the boundary between them, which occurs at the transition voltage, V_{trans}, is indicated with an arrow and exhibits a clear gate dependence. In the low-bias region, the logarithmic dependence is characteristic of direct tunnelling ($V < \Phi_B/e$, where Φ_B is the barrier height and e is the elementary charge), whereas in the high-bias region above V_{trans}, the curves yield a linear relation with a negative slope, indicative of Fowler–Nordheim tunnelling or field emission ($V > \Phi_B/e$). The measurement of V_{trans} allows an experimental determination of the height of the energy barrier associated with the tunnelling transport in molecular junctions, given by the difference between the electrode Fermi energy, E_F, and the energy of the nearest molecular orbital (the highest occupied molecular orbital (HOMO) or the lowest unoccupied molecular orbital (LUMO)[18]; see Supplementary Information for details).

We observe a controllable gate-voltage dependence of V_{trans} in the molecular junction. For Au–ODT–Au junctions, V_{trans} shifts to a lower bias as a more negative V_G is applied (Fig. 1b). To explore this further, we plot the measured V_{trans} values against V_G in Fig. 1c. We find that V_{trans} scales linearly and reversibly as a function of V_G. The slope, $\alpha = \Delta V_{trans}/\Delta V_G$, is the gate efficiency factor, which describes the effectiveness of molecular orbital gating; for example, if $e|\alpha| = 0.25\,\text{eV V}^{-1}$ (from the linear fit in Fig. 1c) then the molecular orbital energy changes by 0.25 eV when 1 V is applied to the gate electrode. This large gate coupling probably indicates that the molecule resides close to the gate metal–dielectric interface to reduce screening by the electrodes; future improvements to the device performance may be made by using a tapered-electrode approach[13]. We can also determine the actual amount of molecular orbital shift produced by the applied gate voltage in terms of an effective molecular orbital gating energy, $eV_{G,\text{eff}} = e|\alpha|V_G$.

[1]Department of Nanobio Materials and Electronics, [2]Department of Materials Science and Engineering, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea. [3]Department of Applied Physics, Hanyang University, Ansan 426-791, South Korea. [4]Departments of Electrical Engineering and Applied Physics, Yale University, New Haven, Connecticut 06520, USA. †Present address: Department of Physics, University of Konstanz, D-78457 Konstanz, Germany.
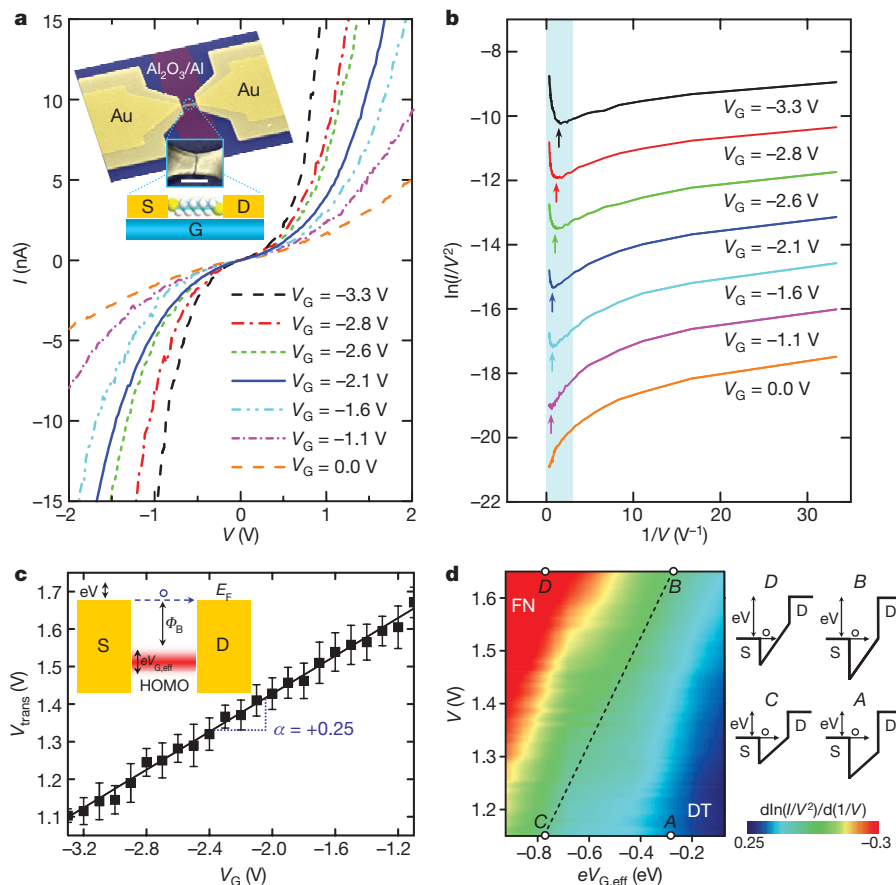
Figure 1 | Gate-controlled charge transport characteristics of a Au-ODT-Au junction. a, Representative $I(V)$ curves measured at 4.2 K for different values of V_G. Inset, the device structure and schematic. S, source; D, drain; G, gate. Scale bar, 100 nm. **b**, Fowler–Nordheim plots corresponding to the $I(V)$ curves in **a**, exhibiting the transition from direct to Fowler–Nordheim tunnelling with a clear gate dependence. The plots are offset vertically for clarity. The arrows indicate the boundaries between transport regimes (corresponding to V_{trans}). **c**, Linear scaling of V_{trans} in terms of V_G. The error bars denote the s.d. of individual measurements for several devices and the solid line represents a linear fit. Inset, the schematic of the energy band for HOMO-mediated hole tunnelling, where $eV_{G,eff}$ describes the actual amount of molecular orbital shift produced by gating. **d**, Two-dimensional colour map of $d\ln(I/V^2)/d(1/V)$ (from Fowler–Nordheim plots). Energy-band diagrams corresponding to four different regions (points A–D) are also shown. FN, Fowler–Nordheim tunnelling; DT, direct tunnelling.

The gate-controlled tunnelling transport is fully illustrated in Fig. 1d, where we show a two-dimensional colour map of $d\ln(I/V^2)/d(1/V)$ (from Fowler–Nordheim plots) in the transition region, as a function of V and $eV_{G,eff}$. The dashed line indicating the linear fit of V_{trans} versus V_G defines the boundary between two distinct transport regimes (direct and Fowler–Nordheim tunnelling). In Fig. 1d, the transport at point A shows typical direct tunnelling across the trapezoidal barrier at given values of V and $eV_{G,eff}$. If the applied bias increases further, to reach V_{trans} (point B), a transition from a trapezoidal to a triangular barrier shape occurs. This indicates the onset of Fowler–Nordheim tunnelling, corresponding to an inflection point in the plot of $\ln(I/V^2)$ against $1/V$. Point C shows V_{trans} to be at a lower bias than at point B, which is attributed to a decrease in the $|E_F - E_{HOMO}|$ offset (where E_{HOMO} is the energy of the HOMO level), arising from a more negative V_G value. Finally, at point D, the barrier is a completely triangular shape and the charge transport is dominated by Fowler–Nordheim tunnelling.

In a three-terminal device, a negative or positive gate voltage would respectively raise or lower the orbital energies in the molecules relative to E_F (refs 1, 6). Hence, a positive value of α indicates HOMO-mediated hole tunnelling (p-type-like; Fig. 1c, inset). Conversely, α would be negative for LUMO-mediated electron tunnelling (n-type-like); for example, a Au–1,4-benzenedicyanide–Au junction (Supplementary Fig. 8). By extrapolating the y intercept from the linear fit in Fig. 1c, we obtain the zero-gate transition voltage, $V_{trans,0} = 1.93 \pm 0.06$ V, which provides an estimate of the original position (at $V_G = 0$ V) of the HOMO level relative to E_F in Au–ODT–Au junctions.

Now let us examine a π-conjugated BDT molecule, the prototypical molecular transport junction[4,6,14]. Figure 2a shows the gate-modulated transport of a Au–BDT–Au junction. The tunnelling current flowing through the BDT junction is enhanced when a negative V_G is applied, whereas a positive V_G suppresses the current level (p-type-like; see also Supplementary Fig. 7). Figure 2b and Fig. 2c respectively display plots of $\ln(I/V^2)$ versus $1/V$ and V_{trans} versus V_G for this molecular system. The positive sign of $\alpha = +0.22$ in the BDT junction explicitly indicates that HOMO-mediated tunnelling is the dominant transport channel. The colour map of $d\ln(I/V^2)/d(1/V)$ (from Fowler–Nordheim plots) is shown as a function of V and $eV_{G,eff}$ (Fig. 2c, inset); the solid line and the dotted arrow indicate the linear fit and the y intercept obtained from the plot of V_{trans} versus V_G, respectively. We find that $V_{trans,0} = 1.14 \pm 0.04$ V for the BDT junction, which is much less than the value for the ODT junction owing to the π-conjugated BDT molecule having a smaller HOMO–LUMO gap.

These results demonstrate relative movement of the molecular orbitals with respect to E_F, accomplished by gating. This should give rise to a modified coupling of the molecular orbitals to the electrodes, depending both on the orbital Fermi level spacing and on the character of the molecular orbital (that is, σ-saturated or π-conjugated). An unambiguous method of investigating this in current-carrying molecular junctions is inelastic electron tunnelling (IET) spectroscopy, which is a powerful tool for investigating the molecular role in charge transport, specifically the coupling between charge carriers and molecular vibrations[3,19–22]. We performed IET spectroscopy
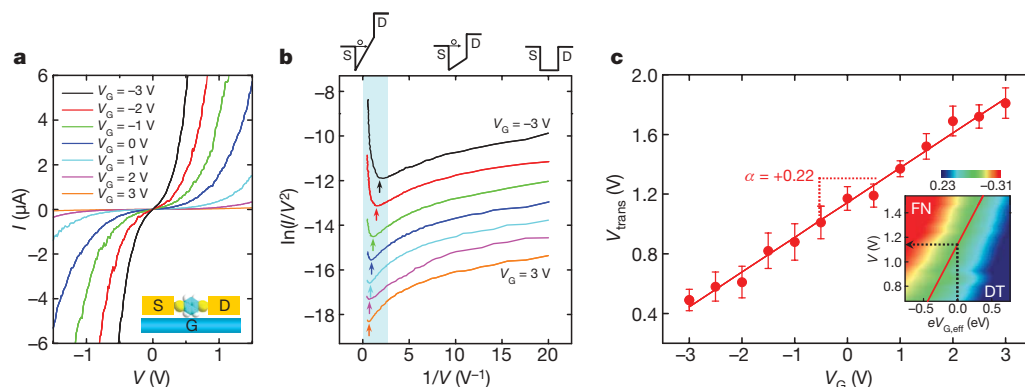
Figure 2 | Gate-controlled charge transport characteristics of a Au-BDT-Au junction. a, Representative $I(V)$ curves measured at 4.2 K for different values of V_G. **b,** Fowler–Nordheim plots demonstrating the gate-variable transition from direct to Fowler–Nordheim tunnelling (colour-coded as in **a**). The plots are offset vertically for clarity. Also shown are drawings of the barrier shape with increasing bias. **c,** Plot of V_{trans} versus V_G. The solid line is a linear fit and the error bars denote the s.d. of the individual measurements. Inset, the colour map of $d\ln(I/V^2)/d(1/V)$ (from Fowler–Nordheim plots) with linear fit (solid line) obtained from the plot of V_{trans} versus V_G. The zero-gate transition voltage is indicated by the dashed arrow.

measurements on the two types of molecules (ODT and BDT). Figures 3a and 4a show the IET spectra of ODT and BDT junctions measured at various values of $eV_{G,eff}$, respectively. A standard a.c. modulation technique with a lock-in amplifier is carried out at 4.2 K to directly acquire the first (dI/dV) and second (d^2I/dV^2) harmonic signals (see Supplementary Information for details). We assign the observed spectral features to specific molecular vibrations by comparison with previously reported infrared, Raman and IET spectroscopy measurements, and also by density functional theory (DFT) calculations (Supplementary Table 1). All of the spectral features are

attributable to vibrational modes associated with the molecular species, which provides additional verification of the constituent molecules.

A well-established validation method for IET spectroscopy is to examine the linewidth broadening as a function of modulation voltage and temperature[20,21]. Figure 3c shows the modulation broadening of a representative IET spectroscopy feature (the 357-mV ν(C–H) stretching mode in the ODT junction) at a constant temperature of 4.2 K. The data points show the FWHM of the experimental peak. Taking into account the known thermal broadening ($5.4k_BT$,
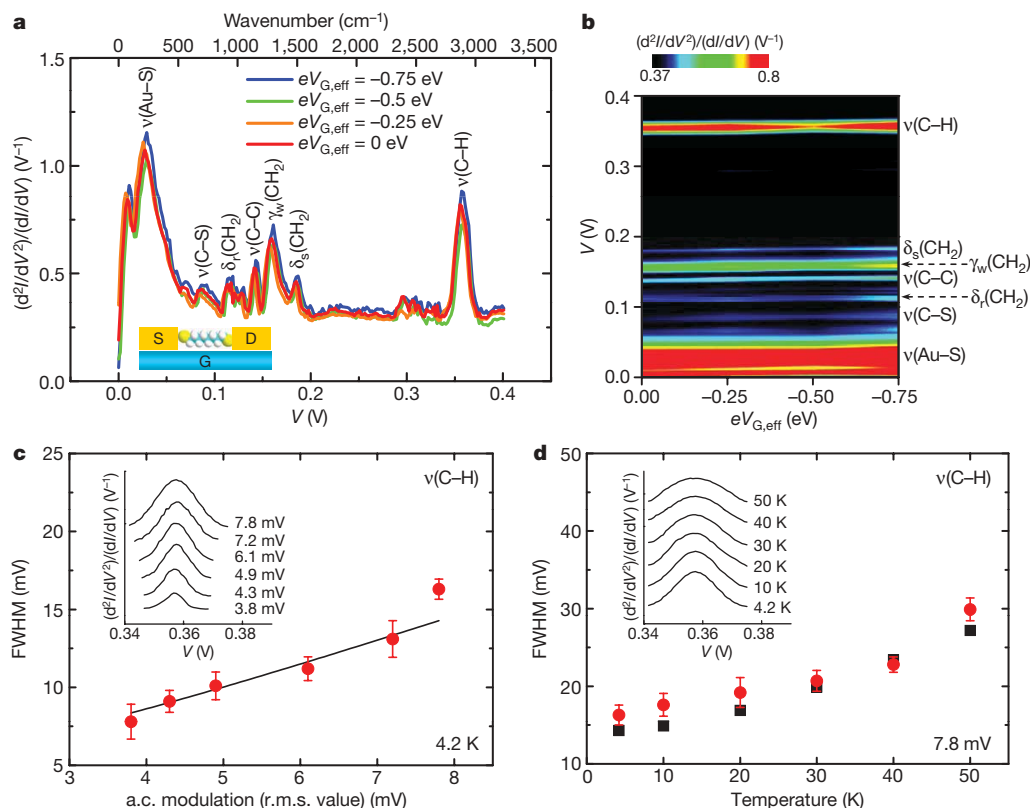

Figure 3 | Gated IET spectra and linewidth broadening of a Au-ODT-Au junction. a, IET spectra measured at 4.2 K for different values of $eV_{G,eff}$, with vibration modes assigned. **b,** Two-dimensional colour map of the gated IET spectra, indicating near independence with respect to $eV_{G,eff}$. **c, d,** Full-width at half-maximum (FWHM) of the peak corresponding to the ν(C–H) stretching mode (~357 mV) as a function of a.c. modulation voltage (**c**) and temperature (**d**). The circles indicate experimental data, and the solid line (**c**) and squares (**d**) show theoretical values. The error bars are determined by the Gaussian fitting. Insets, successive IET spectroscopy scans for the ν(C–H) mode under increasing a.c. modulation voltage (**c**) and increasing temperature (**d**), as indicated. r.m.s., root mean squared.
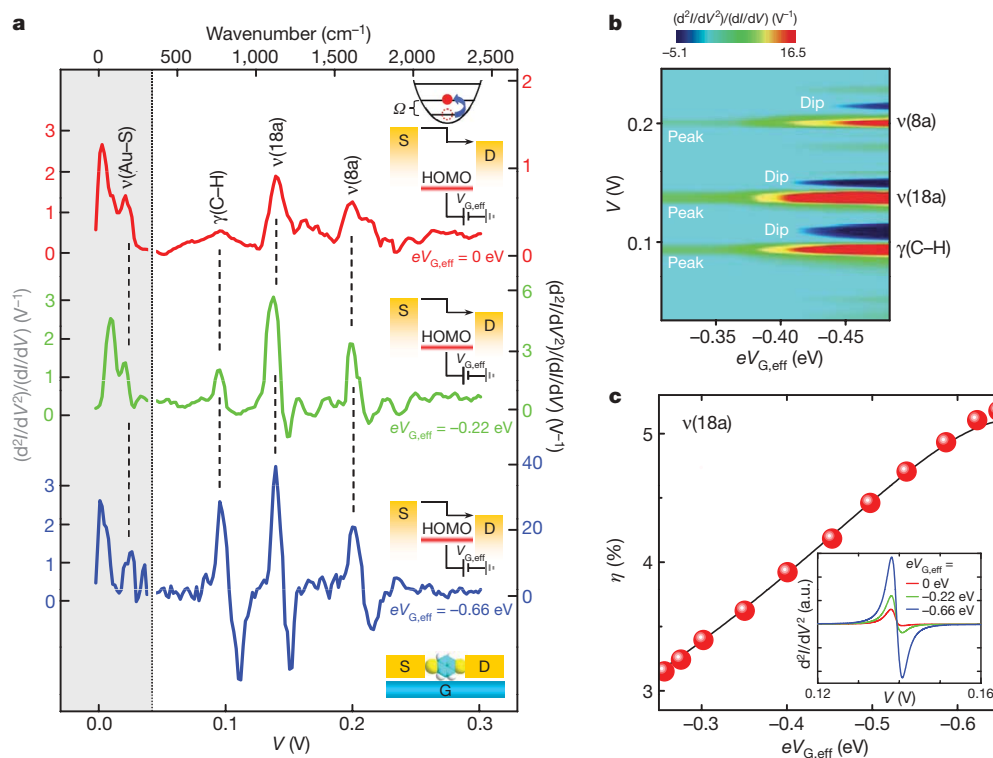
Figure 4 | Resonantly enhanced IET spectra of a Au-BDT-Au junction.
a, IET spectra measured at 4.2 K for different values of $eV_{G,eff}$, with vibration modes assigned. The left-hand y axis corresponds to the grey shaded region of the spectra, and the various right-hand y axes (with different scales) correspond to the related (colour-coded) spectra in the non-shaded region. The vertical dotted line corresponds to $V = 45$ mV (363 cm^{-1}). Significant modification in the spectral intensity and line shape for the benzene ring modes, γ(C–H), ν(18a) and ν(8a), was observed for different values of $eV_{G,eff}$, as indicated. Insets, energy diagrams illustrating inelastic tunnelling as the position of the HOMO resonance shifts as a result of gating. **b**, Two-dimensional colour map of the gated IET spectra, showing that IET spectroscopy intensity and line shape vary significantly as functions of $eV_{G,eff}$. **c**, The relative change, η, in the normalized conductance for the ν(18a) mode as a function of $eV_{G,eff}$. The circles show the experimental data and the solid curve represents the theoretical fit calculated from equation (1). Inset, the gate-variable IET spectra for the ν(18a) mode, simulated using equation (1). a.u., arbitrary units.

where k_B is Boltzmann's constant and T denotes temperature) and modulation broadening ($1.7V_{ac}$, where V_{ac} is the a.c. modulation voltage), the intrinsic linewidth, W_I, can be determined from a fit to the modulation broadening data (Fig. 3c, solid line), giving $W_I = 4.94 \pm 0.89$ meV. Figure 3d shows the thermal broadening of the same ν(C–H) peak at a fixed modulation, demonstrating excellent agreement between experimental FWHM values (circles) and theoretical values (squares) (see Supplementary Fig. 9 for the BDT junction).

We now focus on the gate-voltage dependence of the IET spectra measured in the two types of molecular junction. The spectra of the ODT junction remain essentially unchanged at different values of $eV_{G,eff}$, as shown in Fig. 3a. This constancy with respect to $eV_{G,eff}$ can be also identified in Fig. 3b, which shows a colour map of the gated IET spectra. This result implies that, although the total current is sensitively dependent on the $|E_F - E_{HOMO}|$ barrier height (because it is tunnelling), the inelastic tunnelling features that correspond to vibrational excitations of the molecule are nearly independent of the energetic parameters in the σ-saturated ODT junction, indicating little orbital coupling. However, the IET spectra of the π-conjugated BDT junction are drastically modified by gating (Fig. 4a, b), exhibiting an enhanced spectral intensity and a change in line shape for specific vibrational modes. As $eV_{G,eff}$ becomes more negative in Fig. 4a (that is, as E_{HOMO} is brought closer to E_F), the normalized amplitude $(d^2I/dV^2)/(dI/dV)$ of some IET spectroscopy features in the BDT junction increase in magnitude by a factor of more than 30 and peak shapes of the zero-gate spectrum change to distinct peak-derivative-like (or Fano-shaped) features.

The difference between the Fermi energy and the energy of the molecular orbital has been shown to be an important factor in determining the IET spectroscopy intensity and line shape[23,24]. As illustrated in Figs 1 and 2, the HOMO level for ODT is energetically much farther away from E_F than it is for BDT. Owing to the far-off-resonant nature of the ODT junction, the $eV_{G,eff}$ values observed in the device (Fig. 3a) are predicted not to produce any significant changes in the IET spectra[25]. However, in the opposite limit of near resonance as in the BDT case, the dominant transport orbital (that is, the HOMO) will be strongly coupled to the internal vibrational modes (Fig. 4a, insets), resulting in resonantly enhanced IET spectra[26–28]. In addition, strong coupling between the molecular resonance and the vibrational modes is expected only when the energy states of the molecular orbitals are distributed around the bond directly related to the vibrational modes[28,29], which has been shown to be prominent for π–π orbitals (for example for BDT) and minimal for σ–σ orbitals (for example for ODT)[30]. This is corroborated by DFT calculations showing that the HOMO of BDT has a significant contribution from the benzene ring, whereas there is essentially no contribution from the alkyl chain in the HOMO of ODT (Supplementary Fig. 10). We note that the vibrational modes associated with the benzene ring are the only modes enhanced by molecular orbital gating. The zero-bias features within ~40 mV, including the shoulder features assigned as ν(Au–S) modes, are commonly attributed to phonon interactions in the metal leads[21,22,29], and we observe that this metallic phonon regime is not affected by gating, as shown in Fig. 4a (grey shaded region).

For the enhanced resonant coupling in near-resonance IET, it has been shown that near the excitation threshold, $|eV| = \Omega$, for a molecular vibration, the corresponding change in the total tunnelling conductance (normalized to its elastic background) is[27]

$$\eta = \frac{\delta E^2}{(E_M - \bar{E}_F)^2 + \Gamma/2}\left[\frac{(E_M - \bar{E}_F - \Omega)^2 - (\Gamma/2)^2}{(E_M - \bar{E}_F - \Omega)^2 + (\Gamma/2)^2}\theta(eV - \Omega)\right.$$
$$\left. - \frac{1}{\pi}\frac{(E_M - \bar{E}_F - \Omega)\Gamma}{(E_M - \bar{E}_F - \Omega)^2 + (\Gamma/2)^2}\ln\left|\frac{eV - \Omega}{\Delta}\right|\right] \quad (1)$$

Equation (1) implies that the IET spectroscopy intensity and line shape can be varied, depending on the orbital energy, E_M, and the width, Γ, of the molecular resonance and its coupling, δE, to a particular molecular vibration (where Δ is the cut-off energy, which is much smaller than Γ, and $\bar{E}_F = E_F - eV$). Recently, theoretical models based on the non-equilibrium Green's function formalism have also described such variations in the IET spectroscopy features as the energy of the resonant level shifts[23,24,26]. Using equation (1), we can reproduce the experimental results for the gate-variable IET spectra (see Supplementary Information for details). In Fig. 4c, we show the agreement between the experimental values (data points) and calculated values (solid curve) of η for the ~138-mV ν(18a) stretch mode of the BDT junction as a function of $eV_{G,eff}$, which suggests that equation (1) can provide useful estimates of the contribution from the resonant coupling to the gate-variable IET spectra. We also quantitatively fit the observed IET spectroscopy line shapes for the $eV_{G,eff}$ values shown in Fig. 4a. The agreement between theory and experiment is good for both the spectral line shape and the relative magnitude (Fig. 4c, inset). The same analysis for other vibration modes in the spectra generates essentially similar results to that for the ν(18a) mode.

Our results demonstrate direct gate modulation of molecular orbitals. The IET spectra reveal which orbitals are resonantly enhanced, and dramatic differences are seen in the comparison between near-resonant and far-from-resonant systems. This demonstration validates the concept of molecular-orbital-modulated carrier transport, allowing new designs of molecular-based devices and elucidating both carrier transport mechanisms and the electronic structure of molecular junctions.

METHODS SUMMARY

We used electron-beam lithography to pattern continuous thin gold wires with minimum widths of approximately 100 nm on top of an aluminium gate electrode, previously oxidized in a fresh oxygen atmosphere to form an Al_2O_3 layer typically ~3 nm thick that acts as a gate dielectric. Molecular deposition on the gold surface was performed in a dilute ethanol solution (1 mM) of each molecule inside a nitrogen-filled glove box. Before use, each sample was rinsed in ethanol and gently blown dry in a nitrogen stream. Then we immediately cooled the samples coated with the molecules to 4.2 K (in a vacuum cryostat), and used electromigration to form electrode pairs with a nanometre-scale separation by ramping up a d.c. voltage across the wire while monitoring the resistance. Full details on methods are presented in the Supplementary Information.

1. Ahn, C. H. et al. Electrostatic modification of novel materials. Rev. Mod. Phys. 78, 1185–1212 (2006).
2. Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. Nature 408, 541–548 (2000).
3. Galperin, M., Ratner, M. A., Nitzan, A. & Troisi, A. Nuclear coupling and polarization in molecular transport junctions: beyond tunneling to function. Science 319, 1056–1060 (2008).
4. Reed, M. A., Zhou, C., Muller, C. J., Burgin, T. P. & Tour, J. M. Conductance of a molecular junction. Science 278, 252–254 (1997).
5. Aviram, A. & Ratner, M. A. Molecular rectifiers. Chem. Phys. Lett. 29, 277–283 (1974).
6. Ghosh, A. W., Rakshit, T. & Datta, S. Gating of a molecular transistor: electrostatic and conformational. Nano Lett. 4, 565–568 (2004).
7. Andrews, D. Q., Solomon, G. C., Van Duyne, R. P. & Ratner, M. A. Single molecule electronics: increasing dynamic range and switching speed using cross-conjugated species. J. Am. Chem. Soc. 130, 17309–17319 (2008).
8. Park, J. et al. Coulomb blockade and the Kondo effect in single-atom transistors. Nature 417, 722–725 (2002).
9. Liang, W., Shores, M. P., Bockrath, M., Long, J. R. & Park, H. Kondo resonance in a single-molecule transistor. Nature 417, 725–729 (2002).
10. Kubatkin, S. et al. Single-electron transistor of a single organic molecule with access to several redox states. Nature 425, 698–701 (2003).
11. Xu, B., Xiao, X., Yang, X., Zang, L. & Tao, N. Large gate modulation in the current of a room temperature single molecule transistor. J. Am. Chem. Soc. 127, 2386–2387 (2005).
12. Piva, P. G. et al. Field regulation of single-molecule conductivity by a charged surface atom. Nature 435, 658–661 (2005).
13. Datta, S. S., Strachan, D. R. & Johnson, A. T. C. Gate coupling to nanoscale electronics. Phys. Rev. B 79, 205404 (2009).
14. Ghosh, S. et al. Device structure for electronic transport through individual molecules using nanoelectrodes. Appl. Phys. Lett. 87, 233509 (2005).
15. Xu, B. & Tao, N. J. Measurement of single-molecule resistance by repeated formation of molecular junctions. Science 301, 1221–1223 (2003).
16. Venkataraman, L., Klare, J. E., Nuckolls, C., Hybertsen, M. S. & Steigerwald, M. L. Dependence of single-molecule junction conductance on molecular conformation. Nature 442, 904–907 (2006).
17. Wang, W., Lee, T. & Reed, M. A. Mechanism of electron conduction in self-assembled alkanethiol monolayer devices. Phys. Rev. B 68, 035416 (2003).
18. Beebe, J. M., Kim, B., Gadzuk, J. W., Frisbie, C. D. & Kushmerick, J. G. Transition from direct tunneling to field emission in metal-molecule-metal junctions. Phys. Rev. Lett. 97, 026801 (2006).
19. Jaklevic, R. C. & Lambe, J. Molecular vibration spectra by electron tunneling. Phys. Rev. Lett. 17, 1139–1140 (1966).
20. Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. Science 280, 1732–1735 (1998).
21. Wang, W., Lee, T., Kretzschmar, I. & Reed, M. A. Inelastic electron tunneling spectroscopy of an alkanedithiol self-assembled monolayer. Nano Lett. 4, 643–646 (2004).
22. Kushmerick, J. G. et al. Vibronic contributions to charge transport across molecular junctions. Nano Lett. 4, 639–642 (2004).
23. Galperin, M., Ratner, M. A. & Nitzan, A. On the line widths of vibrational features in inelastic electron tunneling spectroscopy. Nano Lett. 4, 1605–1611 (2004).
24. Shimazaki, T. & Asai, Y. Theoretical study of the lineshape of inelastic electron tunneling spectroscopy. Phys. Rev. B 77, 115428 (2008).
25. Troisi, A., Ratner, M. A. & Nitzan, A. Vibronic effects in off-resonant molecular wire conduction. J. Chem. Phys. 118, 6072–6082 (2003).
26. Galperin, M., Ratner, M. A. & Nitzan, A. Inelastic electron tunneling spectroscopy in molecular junctions: peaks and dips. J. Chem. Phys. 121, 11965–11979 (2004).
27. Persson, B. N. J. & Baratoff, A. Inelastic electron tunneling from a metal tip: the contribution from resonant processes. Phys. Rev. Lett. 59, 339–342 (1987).
28. Komeda, T. Chemical identification and manipulation of molecules via vibrational excitation through inelastic tunneling process with scanning tunneling microscopy. Prog. Surf. Sci. 78, 41–85 (2005).
29. Yu, L. H., Zangmeister, C. D. & Kushmerick, J. G. Origin of discrepancies in inelastic electron tunneling spectra of molecular junctions. Phys. Rev. Lett. 98, 206803 (2007).
30. Troisi, A. & Ratner, M. A. Propensity rules for inelastic electron tunneling spectroscopy of single-molecule transport junctions. J. Chem. Phys. 125, 214709 (2006).

Author Contributions T.L. planned and supervised the project; H.S. designed and performed the experiments; H.S., T.L. and M.A.R. analysed and interpreted the data and wrote the manuscript; H.J. designed the electrical measurement systems; Y.K. assisted in low-temperature electrical measurements; and Y.H.J. performed DFT calculations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.L. (tlee@gist.ac.kr) or M.A.R. (mark.reed@yale.edu).

LETTERS

Glaciers as a source of ancient and labile organic matter to the marine environment

Eran Hood[1], Jason Fellman[2]†, Robert G. M. Spencer[3]†, Peter J. Hernes[3], Rick Edwards[4], David D'Amore[4] & Durelle Scott[5]

Riverine organic matter supports of the order of one-fifth of estuarine metabolism[1]. Coastal ecosystems are therefore sensitive to alteration of both the quantity and lability of terrigenous dissolved organic matter (DOM) delivered by rivers. The lability of DOM is thought to vary with age, with younger, relatively unaltered organic matter being more easily metabolized by aquatic heterotrophs than older, heavily modified material[2–4]. This view is developed exclusively from work in watersheds where terrestrial plant and soil sources dominate streamwater DOM. Here we characterize streamwater DOM from 11 coastal watersheds on the Gulf of Alaska that vary widely in glacier coverage (0–64 per cent). In contrast to non-glacial rivers, we find that the bioavailability of DOM to marine microorganisms is significantly correlated with increasing ^{14}C age. Moreover, the most heavily glaciated watersheds are the source of the oldest (~4 kyr ^{14}C age) and most labile (66 per cent bioavailable) DOM. These glacial watersheds have extreme runoff rates, in part because they are subject to some of the highest rates of glacier volume loss on Earth[5]. We estimate the cumulative flux of dissolved organic carbon derived from glaciers contributing runoff to the Gulf of Alaska at $0.13 \pm 0.01\,\mathrm{Tg\,yr^{-1}}$ ($1\,\mathrm{Tg} = 10^{12}\,\mathrm{g}$), of which ~0.10 Tg is highly labile. This indicates that glacial runoff is a quantitatively important source of labile reduced carbon to marine ecosystems. Moreover, because glaciers and ice sheets represent the second largest reservoir of water in the global hydrologic system, our findings indicate that climatically driven changes in glacier volume could alter the age, quantity and reactivity of DOM entering coastal oceans.

Biogeochemical cycling in coastal margins near riverine outflows is dominated by the influx of terrestrial organic matter and nutrients. The effect of anthropogenic increases in nutrient export on sensitive systems is well-documented in regions such as the Gulf of Mexico zone of hypoxia[6]. It is much less clear how climate-induced shifts in the export of terrigenous DOM will affect coastal environments, although the reactivity of this carbon will be key as the extent of its incorporation into marine food webs depends largely on its chemical character[7]. Riverine DOM is typically dominated by allochthonous material derived from plant detritus, which may be substantially aged and degraded before entering the marine environment, thereby reducing its availability to marine heterotrophs[3]. Glacial ecosystems are devoid of higher plants but contain abundant microbial communities adapted to extreme temperature, light and nutrient regimes[8]. Thus, we hypothesize that glacial watersheds export chemically distinct DOM compared to DOM derived from litter and soil organic matter in forested watersheds.

The Gulf of Alaska (GOA) drainage basin contains more than 10% of the mountain glaciers on Earth[5], and annual runoff from this region represents the second largest discharge of freshwater to the

Pacific Ocean[9,10]. Along the GOA, watersheds dominated by glaciers have extremely high water fluxes (commonly $>6\,\mathrm{m\,yr^{-1}}$), such that yields of DOM from these watersheds are substantial[11] even with the low concentrations of DOM typical of glacial ecosystems[12,13]. Future changes in glacial runoff are markedly larger than those projected for other components of the water cycle, and the continued recession of GOA glaciers with climate warming is expected to shift the timing and magnitude of riverine DOM delivery to downstream coastal ecosystems[11].

In this study we characterized the source, bioavailability and magnitude of the DOM flux from glacial watersheds that drain into the GOA. We sampled streamwater DOM from 11 coastal watersheds in three geographic areas along the GOA during peak glacial runoff (Fig. 1). All three study areas are located in coastal temperate rainforest biomes characterized by maritime climates, with mean annual sea level temperatures ranging from 2.5 to 4.5 °C. The 11 watersheds are largely free of human disturbance, and land cover is dominated by coniferous forest in the lower elevations of all of the watersheds. Glacier coverage varied markedly (0–64%) within the watersheds owing to differences in elevation and distance from the ocean. Here we use glacier cover as a relative estimate of the proportion of streamwater derived from glacial runoff in each watershed. All riverine water samples were collected above tidal influence within 6 km of tidewater. In addition to measuring bulk dissolved organic carbon (DOC), we used fluorescence spectroscopy, the vascular plant biomarker lignin, and carbon isotopic analyses (δ^{13}C and Δ^{14}C) to evaluate the chemical character, source and age of riverine DOM exported from the 11 watersheds.

Streamwater DOC concentrations ranged from 0.6 to 2.2 mg Cl^{-1} and were negatively correlated with watershed glacier coverage (Table 1; $r^2 = 0.63$, $P = 0.003$). This is consistent with higher inputs of soluble carbon from forest soils in the watersheds with lower glacier coverage[13]. Fluorescence spectroscopy and δ^{13}C-DOC measurements both indicated that increased inputs of glacial meltwater were associated with an increase in the proportion of DOM derived from microbial sources. DOM fluorescence associated with proteinaceous material was positively correlated with watershed glacier coverage (Table 1; $r^2 = 0.82$, $P = 0.001$), indicating that DOM in heavily glaciated watersheds has a high protein content and relatively low concentrations of aromatic organic compounds that typically dominate streamwater DOM[14]. The δ^{13}C values for streamwater DOC ranged from $-25.6‰$ to $-22.0‰$ and similarly demonstrated significant enrichment with increasing watershed glacier coverage (Table 1; $r^2 = 0.53$, $P = 0.01$). The δ^{13}C-DOC values in the heavily ($>40\%$) glacial watersheds were all enriched compared to the range for C$_3$ plant-derived carbon (about -25 to $-30‰$)[3,15], which is
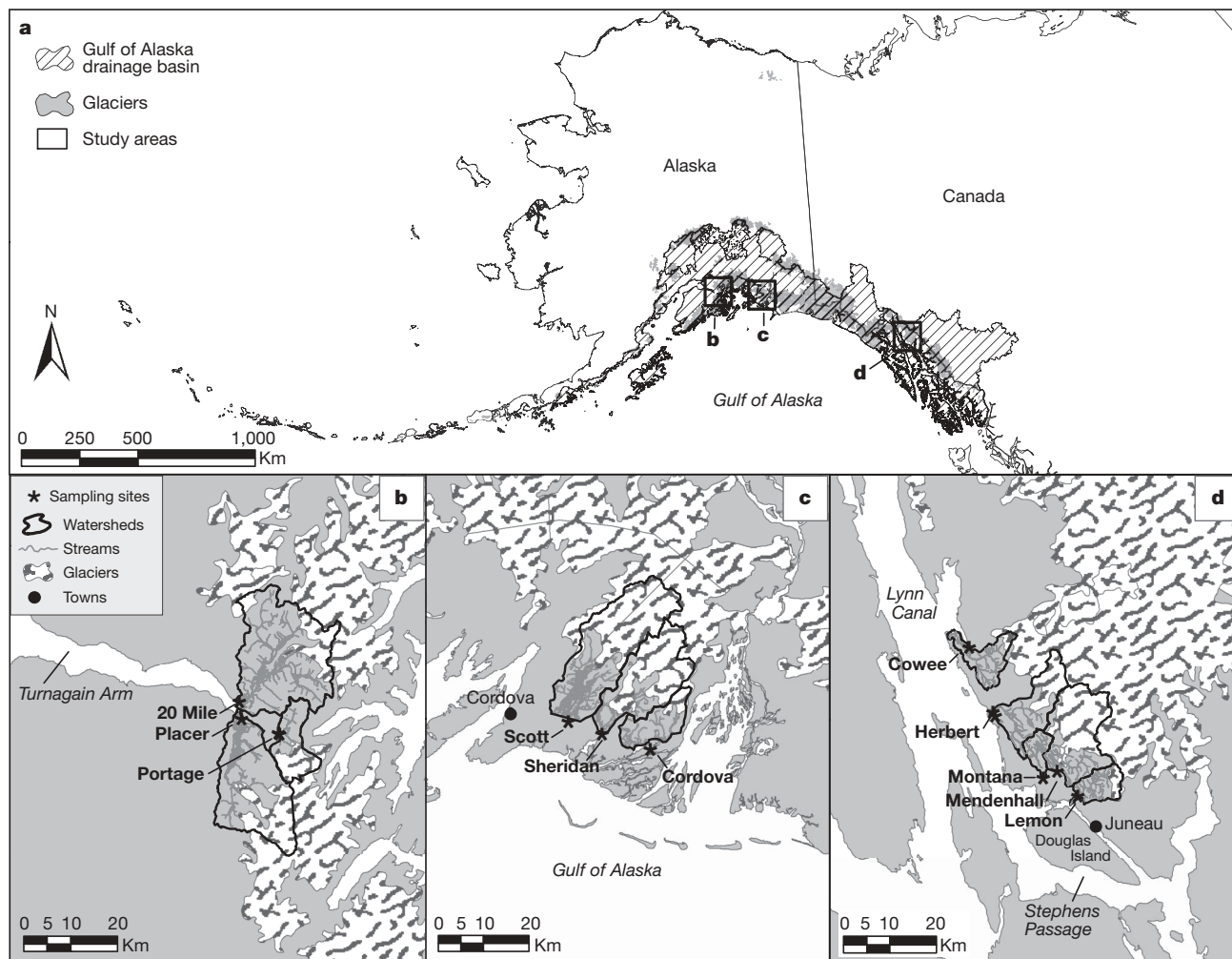
Figure 1 | Map of the Gulf of Alaska drainage area. a, The region draining into the GOA covers 420,500 km^2 within the United States and Canada and contains 75,300 km^2 of glaciers (shown in dark grey). **b–d,** Eleven watersheds (named in bold font) around the GOA were sampled within 6 km of their estuaries in three locations: Girdwood (**b**), Cordova (**c**) and Juneau (**d**). The watershed areas ranged from 37 to 430 km^2 (mean, 201; s.e., 41).

consistent with DOC derived from autochthonous material in fresh-water and marine ecosystems[16,17].

Unlike δ^{13}C, which is a bulk property of DOM, lignin phenols are a unique biomarker for vascular plants and thus terrigenous plant material in aquatic ecosystems[17,18]. Across the 11 watersheds, carbon-normalized lignin yields decreased dramatically with increasing water-shed glacier area (Table 1; $r^2 = 0.91$, $P < 0.001$). This finding indicates that the bulk of the DOM in glacially dominated rivers is not directly derived from higher plants. The presence of abundant, active microbial communities in supra[19]-, sub[8]- and pro[20]-glacial environments,

combined with limited spectroscopic observations[12] and the low C:N values[11] of DOM in glacial streams, have previously led to the hypo-thesis that DOM in glacial runoff has a microbial source. Our findings provide direct corroboration of this hypothesis through three inde-pendent lines of evidence, indicating that increasing glacier contribu-tions to streamflow shift riverine DOM sources from vascular plant material toward microbial biomass.

In order to better understand the fate and impact of glacier-derived microbial DOM on coastal ecosystems, we conducted short-term (2 week) bioavailability incubations of riverine DOC with

Table 1 | Data from sample watersheds

Watershed	Glacier cover (%)	DOC (mg C l^{-1})	Protein fluor. (%)	δ^{13}C-DOC (‰)	Lignin (mg per 100 mg OC)	Bioavailable DOC (%)
Montana	0	2.2	2.7	−25.3	0.45	23
Cowee	9	1.5	4.2	−24.4	0.38	26
Cordova	16	1.4	10.1	−25.6	0.36	36
Lemon	25	1.1	23.7	−23.1	0.23	42
20 Mile	37	1.7	20.3	−24.1	0.14	31
Herbert	38	1.1	25.2	−24.5	0.25	40
Placer	48	1.4	33.9	−22.6	0.10	60
Portage	52	0.8	41.1	−24.6	0.11	48
Scott	52	1.0	55.5	−22.9	0.08	64
Mendenhall	55	0.7	39.6	−23.4	0.11	41
Sheridan	64	0.6	77.0	−22.0	0.09	66

Shown are glacier coverage, concentrations of DOC, and chemical and spectroscopic characteristics of DOM. Protein fluorescence is the relative contribution of the two DOM fluorescence components associated with proteinaceous material to the total fluorescence of the nine fluorescence components identified by our PARAFAC model as described in Methods. δ^{13}C values are expressed as per mil relative to the Pee Dee Belemnite standard, where δ^{13}C = [($R_{sample}/R_{standard}$) − 1] × 1,000, and R is the ratio of ^{13}C to ^{12}C; replicate analyses yielded δ^{13}C-DOC values with standard deviations of ±0.1‰. Lignin carbon-normalized yields are reported as the sum of three syringyl, three vanillyl and two cinnamyl phenols.

inoculums of microorganisms from downstream near-shore marine ecosystems. DOM derived from microbially based ecosystems such as Antarctic dry lakes that lack inputs from higher plants is thought to be highly labile because of low C:N ratios and low aromatic carbon content[21]. We found that the lability of riverine DOC increased dramatically with glacier coverage (Table 1), and that the bioavailability of DOC was highly correlated with Δ^{14}C-DOC values across the 11 watersheds (Fig. 2). In the most heavily glaciated watershed, Sheridan River, 66% of the riverine DOC was readily degraded by marine microbes despite having a Δ^{14}C value of -386‰ ($\sim$3,900 yr Δ^{14}C age). Heterotrophic microbes in both sub-glacial[22] and pro-glacial[20] environments have been shown to subsist on aged carbon overrun by ice during periods of glacier advance. It is additionally possible that CO_2 respired from glacially sequestered carbon may support microbial primary production in glacial ecosystems. Along the GOA, the last major cycle of glacier retreat and re-advance occurred during the Hypsithermal warm period between 7,000 and 2,500 yr BP[23]. Our data indicate that microorganisms in GOA glacial ecosystems can liberate carbon fixed during the Hypsithermal as DOC in a form that is highly bioavailable to aquatic heterotrophs. Riverine DOC in the less glacial watersheds was substantially less bioavailable despite having modern, enriched Δ^{14}C values consistent with inputs of recently fixed carbon from temperate forest soils. These findings contradict the prevalent view that the labile fraction of riverine DOC is dominated by young, lightly degraded material[2–4]. Moreover, the fact that glaciers can release ancient DOC has important implications for the interpretation of DOC age in rivers[4] and coastal oceans[24] affected by glacial runoff.

The impact of labile glacially derived DOM on heterotrophic production in GOA marine ecosystems depends on the magnitude of DOM fluxes from glaciers. A GIS (geographic information system) model of annual precipitation on the 75,300 km^2 of glaciers that drain to the GOA was used in combination with previous estimates of distributed glacier volume loss across the region[5] to calculate the annual water flux from GOA glaciers. We estimate specific discharge from GOA glaciers at 5.4 ± 0.6 m yr^{-1} (Table 2), which yields a total annual runoff of 410 ± 40 km^3 yr^{-1} that is comparable to the annual discharge of the Mississippi River[9]. Based on previous measurements of DOC concentrations in glacial runoff[12–14], we suggest that direct runoff from GOA glaciers produces a conservative DOC flux of 0.13 ± 0.01 Tg yr^{-1} to downstream ecosystems (Table 2). Furthermore, the DOC bioavailability value from the most heavily glaciated catchment (Sheridan; Table 1) suggests that $\sim$0.1 Tg of the annual DOC derived from

Table 2 | Total and area-weighted annual fluxes from glaciers draining into the GOA

	Water		DOC	
	(km^3 yr^{-1})	(m yr^{-1})	(10^9 kg yr^{-1})	(kg km^{-2} yr^{-1})
Glacial runoff	320 ± 30	4.2 ± 0.4	0.10 ± 0.01	$1,300 \pm 110$
Glacial wastage	90 ± 30	1.2 ± 0.4	0.03 ± 0.01	350 ± 120
Total flux	410 ± 30	5.4 ± 0.6	0.13 ± 0.01	$1,650 \pm 160$

Glacial runoff is the annual material flux from glaciers assuming glacier volume remains constant. Glacial wastage is the annual material flux resulting from lost glacier volume due to thinning and retreat (from ref. 5 for the period mid-1990s to 2001). The wastage of GOA glaciers generates approximately 20% of the total glacier fluxes of water and carbon.

GOA glaciers is readily bioavailable. Because glacial streamwater turbidities are high and riverine transit times from glaciers to their estuaries are short across broad regions of the GOA, a substantial portion of this labile DOC is probably delivered to marine heterotrophic communities without biological or photochemical alteration. The biogeochemical significance of inputs of labile DOM discharged from glaciers is magnified by the relatively small volume of the GOA. On a per volume basis, the flux of labile glacier-derived DOM to the shelf waters of the GOA is approximately 25–50% of the total labile DOM input to the polar surface waters of the Arctic Ocean[25,26], which has the highest per volume inputs of DOM of any ocean on Earth[27].

The quantitative importance of glacial ecosystems as sources of riverine DOM has been greatly underappreciated, as evidenced by the dearth of studies on DOM dynamics in glacial rivers[12,13]. However, the annual area-weighted flux of DOC that we report for GOA glaciers (1,650 kg km^{-2} yr^{-1}; Table 2) is higher than the average per area flux of DOC from the watersheds of the five largest rivers draining to the Arctic Ocean (1,600 kg km^{-2} yr^{-1})[4]. Thus, we conclude that glacial ecosystems, which cover approximately 15,000,000 km^2 or $\sim$10% of the Earth's surface, have the potential for substantial DOM export in a form that is highly available to aquatic heterotrophs. Along the GOA, glacier wastage accounts for a substantial portion ($\sim$20%, Table 2) of the total riverine flux of water and DOM. As a result, fluxes of glacier-derived DOM to the GOA are likely to increase in the near term as ice thinning rates[28] and discharge in glacial rivers[29] continue to rise.

Globally, the rate of glacier ice loss from terrestrial storage is increasing, with important implications for sea level[30] and the physical oceanography of coastal oceans[10]. Our findings indicate that future changes in discharge from glacial rivers may have previously unrecognized impacts on coastal biogeochemistry. Because bacterial production in coastal river plumes can be largely supported by terrigenous DOM[31], our results suggest that changes in the magnitude and timing of glacial runoff to the ocean could alter carbon availability and heterotrophic productivity in marine ecosystems. These changes could be particularly pronounced in regions that currently support commercially important fisheries such as the GOA and the North Atlantic, the latter influenced by a 41% increase in Greenland ice sheet discharge between 1961–1990 and 1998–2003[32].

Extending our estimates of glacier-derived DOM export to other regions of the world will require sampling and characterization of DOM in additional glacial rivers. Moreover, the amount and quality of organic matter within and beneath other coastal glaciers may differ from that of the coastal temperate rainforests of Alaska. Even so, our results demonstrate that glacial ecosystems are quantitatively important and highly dynamic sources of reduced carbon to freshwater and marine ecosystems that receive their runoff. The existence of such large and highly labile carbon fluxes from GOA glaciers underscores the uncertainties in carbon budgets of coastal regions, which play critical roles in moderating the flux of reactive species to the ocean.

METHODS SUMMARY

Stream samples from 11 coastal watersheds on the Gulf of Alaska were collected over a three day period in late July 2008 and were filtered within one day of
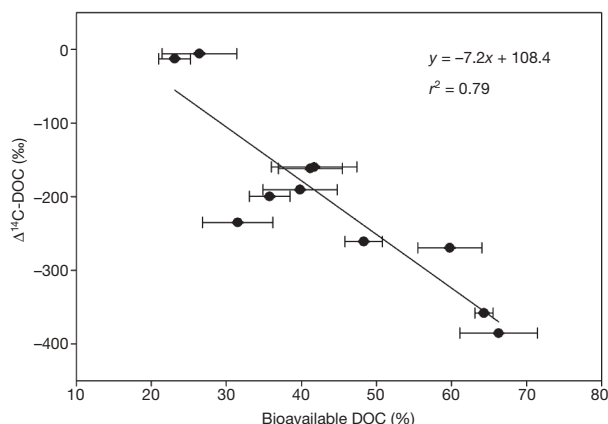

Figure 2 | Relationship between Δ^{14}C-DOC and bioavailable DOC for the 11 rivers sampled. The percentage of riverine DOC that was readily bioavailable to marine microorganisms increased significantly as Δ^{14}C-DOC values became more depleted ($y = -7.2x + 108.4$; $r^2 = 0.79$; $P = 0.001$). Values of Δ^{14}C are expressed as the deviation in per mil (‰) from the ^{14}C activity of nineteenth century wood. Errors ($\pm$1 s.d.) associated with Δ^{14}C AMS analyses averaged 2.0‰ ($\pm$20 yr for radiocarbon age) and error bars are smaller than the symbols. Error bars on bioavailable DOC are 1 s.e.

collection. Water samples were immediately analysed for bulk DOC concentration and DOM fluorescence. Protein fluorescence components were quantified from fluorescence excitation-emission matrices using the multivariate modelling technique parallel factor analysis (PARAFAC), which decomposes the fluorescence spectra of DOM into independent components. Streamwater DOM was also analysed for δ^{13}C-DOC. Lyophilized DOM samples from each site were analysed for lignin phenols and Δ^{14}C-DOC. To determine the bioavailability of DOC from the study streams, a laboratory incubation was conducted using near-shore marine water from the GOA as a microbial inoculum.

Fluxes of water from glaciers along the GOA were estimated using a combination of: (1) runoff derived from PRISM (parameter-elevation regression on independent slopes model) precipitation models that were calibrated with regional streamflow data and (2) runoff derived from the estimates of glacier wastage in ref. 5. Estimates of glacial runoff (1) and glacier wastage (2) were constrained using the regional glacier area from ref. 5 that drains to the GOA. Fluxes of DOC delivered from glaciers to downstream aquatic ecosystems were calculated using a conservative estimate of the average melt season concentration of DOC in glacial runoff in conjunction with the modelled water fluxes from glaciers along the GOA.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Smith, S. V. & Hollibaugh, J. T. Coastal metabolism and the oceanic organic carbon balance. *Rev. Geophys.* **31**, 75–89 (1993).
2. Mayorga, E. *et al.* Young organic matter as a source of carbon dioxide outgassing from Amazonian rivers. *Nature* **436**, 538–541 (2005).
3. Raymond, P. A. & Bauer, J. E. Riverine export of aged terrestrial organic matter to the North Atlantic Ocean. *Nature* **409**, 497–500 (2001).
4. Raymond, P. A. *et al.* Flux and age of dissolved organic carbon exported to the Arctic Ocean: a carbon isotopic study of the five largest arctic rivers. *Glob. Biogeochem. Cycles* **21**, GB4011, doi:10.1029/2007GB002934 (2007).
5. Arendt, A. A., Echelmeyer, K. A., Harrison, W. D., Lingle, C. S. & Valentine, V. B. Rapid wastage of Alaska glaciers and their contribution to rising sea level. *Science* **297**, 382–386 (2002).
6. Rabalais, N. N., Turner, R. E. & Wiseman, W. J. Gulf of Mexico hypoxia, aka ''The dead zone''. *Annu. Rev. Ecol. Syst.* **33**, 235–263 (2002).
7. Sondergaard, M., Stedmon, C. A. & Borch, N. H. Fate of terrigenous dissolved organic matter in estuaries: aggregation and bioavailability. *Ophelia* **57**, 161–176 (2003).
8. Skidmore, M. L., Foght, J. M. & Sharp, M. J. Microbial life beneath a high arctic glacier. *Appl. Environ. Microbiol.* **66**, 3124–3220 (2000).
9. Ludwig, W., Probst, J. L. & Kempe, S. Predicting the oceanic input of organic carbon by continental erosion. *Glob. Biogeochem. Cycles* **10**, 23–41 (1996).
10. Royer, T. C. & Grosch, C. E. Ocean warming and freshening in the northern Gulf of Alaska. *Geophys. Res. Lett.* **33**, L16605, doi:10.1029/2006GL026767 (2006).
11. Hood, E. & Scott, D. Riverine organic matter and nutrients in southeast Alaska affected by glacial coverage. *Nature Geosci.* **1**, 583–587 (2008).
12. Barker, J. D., Sharp, M. J., Fitzsimons, S. J. & Turner, R. J. Abundance and dynamics of dissolved organic carbon in glacier systems. *J. Arct. Antarct. Alpine Res.* **38**, 163–172 (2006).
13. Hood, E. & Berner, L. The effect of changing glacial coverage on the physical and biogeochemical properties of coastal streams in southeastern Alaska. *J. Geophys. Res.* **114**, G03001, doi:10.1029/2009JG000971 (2009).
14. Weishaar, J. L., Aiken, G. R., Bergamaschi, B. A., Fram, M. S. & Fujil, R. Evaluation of specific ultraviolet absorbance as an indicator of the chemical composition and reactivity of dissolved organic carbon. *Environ. Sci. Technol.* **37**, 4702–4708 (2003).
15. Schiff, S. L., Aravena, R., Trumbore, S. E. & Hinton, M. J. Export of DOC from forested catchments on the Precambrian Shield of Central Ontario: clues from ^{13}C and ^{14}C. *Biogeochemistry* **36**, 43–65 (1997).
16. McKnight, D. M., Hood, E. & Klapper, L. in *Aquatic Ecosystems: Interactivity of Dissolved Organic Matter* (eds Findlay, S. E. G. & Sinsabaugh, R. L.) Ch. 3, 71–96 (Academic, 2003).
17. Opsahl, S. & Benner, R. Distribution and cycling of terrigenous dissolved organic matter in the ocean. *Nature* **386**, 480–482 (1997).
18. Hernes, P. J. & Benner, R. Photochemical and microbial degradation of dissolved lignin phenols: implications for the fate of terrigenous dissolved organic matter in marine environments. *J. Geophys. Res.* **108**, 3291, doi:10.1029/2002JC001421 (2003).
19. Anesio, A. M., Hodson, A. J., Fritz, A., Psenner, R. & Sattler, B. High microbial activity on glaciers: importance to the global carbon cycle. *Glob. Change Biol.* **15**, 955–960 (2008).
20. Bardgett, R. D. *et al.* Heterotrophic microbial communities use ancient carbon following glacial retreat. *Biol. Lett.* **3**, 487–490 (2007).
21. McKnight, D. M., Andrews, E. D., Spaulding, S. A. & Aiken, G. R. Aquatic fulvic acids in algal-rich Antarctic ponds. *Limnol. Oceanogr.* **39**, 1972–1979 (1994).
22. Hodson, A. *et al.* Glacial ecosystems. *Ecol. Monogr.* **78**, 41–67 (2008).
23. Deevey, E. S. & Flint, R. F. Postglacial hypsithermal interval. *Science* **125**, 182–184 (1957).
24. Druffel, E. R. M. & Bauer, J. E. Radiocarbon distributions in Southern Ocean dissolved and particulate organic matter. *Geophys. Res. Lett.* **27**, 1495–1498 (2000).
25. Holmes, R. M. *et al.* Lability of DOC transported by Alaskan rivers to the Arctic Ocean. *Geophys. Res. Lett.* **35**, L03402, doi:10.1029/2007GL032837 (2008).
26. Spencer, R. G. M. *et al.* Utilizing chromophoric dissolved organic matter measurements to derive export and reactivity of dissolved organic carbon exported to the Arctic Ocean: a case study of the Yukon River. *Geophys. Res. Lett.* **36**, L060401, doi:10.1029/2008GL036831 (2009).
27. Opsahl, S., Benner, R. & Amon, R. Major flux of terrigenous dissolved organic matter through the Arctic Ocean. *Limnol. Oceanogr.* **44**, 2017–2023 (1999).
28. Larsen, C. F., Motyka, R. J., Arendt, A. A., Echelmeyer, K. A. & Geissler, P. E. Glacier changes in southeast Alaska and northwest British Columbia and contribution to sea level rise. *J. Geophys. Res.* **112**, F01007, doi:10.1029/2006JF000586 (2007).
29. Neal, E. G., Walter, M. T. & Coffeen, C. Linking the Pacific Decadal Oscillation to seasonal stream discharge patterns in southeast Alaska. *J. Hydrol.* **263**, 188–197 (2002).
30. Meier, M. F. *et al.* Glaciers dominate eustatic sea-level rise in the 21st century. *Science* **317**, 1064–1067 (2007).
31. Chin-Leo, G. & Benner, R. Enhanced bacterioplankton production and respiration at intermediate salinities in the Mississippi River plume. *Mar. Ecol. Prog. Ser.* **87**, 87–103 (1992).
32. Hanna, E. *et al.* Runoff and mass balance of the Greenland ice sheet: 1958–2003. *J. Geophys. Res. D* **110**, D13108, doi:10.1029/2004JD005641 (2005).

Author Contributions E.H., J.F., R.G.M.S. and D.S. designed the research. J.F. performed the fieldwork and J.F., R.G.M.S. and E.H. performed laboratory analyses. P.J.H., R.E. and D.D. supported analyses and aided data interpretation. E.H. wrote the paper with substantial contributions from J.F., R.G.M.S. and P.J.H. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.H. (eran.hood@uas.alaska.edu).

nature

METHODS

Sample collection and analysis. Stream samples from each of the 11 sites were collected as composites of three grab samples collected in a well-mixed region of the stream. Samples were stored on ice and returned to the University of Alaska Southeast for processing within one day of collection. Samples were filtered twice with precombusted 47-mm glass fibre filters (0.7 μm followed by 0.3 μm) to remove fine glacial silt and analysed for DOC (Shimadzu TOC-V high-temperature combustion) and DOM fluorescence (Jobin Yvon Horiba Fluoromax-3) at the USDA Forestry Science Lab in Juneau, Alaska. Fluorescence excitation-emission matrices (EEMs) were created by measuring fluorescence intensity across a range of excitation (240–450 nm) and emission (300–600 nm) wavelengths. Parallel factor analysis (PARAFAC) modelling of fluorescence EEMs was conducted with Matlab[33]. PARAFAC was used to identify and determine the relative contribution of protein fluorescence components of DOM. δ^{13}C-DOC samples were analysed at UC Davis using a Model 1010 TOC Analyser (OI Analytical) interfaced to a PDZ Europa 20-20 isotope ratio mass spectrometer (Sercon Ltd). Streamwater DOC from filtered samples was lyophilized using a Labconco freeze drier (FreezeZone 2.5) at $-50\,^{\circ}$C and was stored in a desiccator until analysis. Lignin phenols were analysed at UC Davis as described in ref. 18. Blank concentrations never exceeded 5% of the total lignin phenols in an individual extract. Δ^{14}C-DOC was measured at the UC Irvine KCCAMS facility and values were corrected for sample δ^{13}C. The bioavailability of DOC was determined as the difference in DOC before and after a 14-day laboratory incubation as described in ref. 33. The microbial inoculum was prepared using GOA near-shore marine water filtered through pre-combusted Whatman GF/D filter (nominal pore size 2.7 μm).

Glacial water and carbon flux estimates. Annual runoff from GOA glaciers was estimated as the sum of (1) glacier runoff, the annual water flux from glaciers assuming glacier volume remains constant, and (2) glacier wastage, the annual material flux resulting from lost glacier volume due to thinning and retreat. Glacier runoff was estimated using a GIS layer of glacier coverage in Alaska and the Yukon from ref. 5 in combination with PRISM precipitation models for Alaska, the Yukon Territory, and British Columbia (http://www.prism.oregonstate.edu/). The PRISM data are annual average precipitation data for the period 1960–1990 and are gridded at 2 km^2 (ref. 34). Precipitation data were applied to the glacier area draining to the GOA to derive runoff from glaciers. Annual precipitation on glaciers was converted to annual runoff by calibration against annual discharge from 36 watersheds that contain USGS stream gauges with more than 10 years of records. The GOA drainage basin was divided into two regions. Twenty-one USGS stream gauges in south-central and southwest Alaska (region 1) and fifteen USGS stream gauges along the central GOA coast

and southeastern Alaska (region 2) were used to compute calibration factors that converted annual PRISM precipitation volumes into annual runoff volumes from glaciers in each of the two regions. The majority of the USGS gauges used to compute calibration factors were located in watersheds that contained glaciers. Because glacial wastage was calculated separately (see below), we used watershed glacier area and glacier thinning rates in ref. 5 to calculate the mean annual volume of discharge from glacier wastage in each of the calibration watersheds. Runoff from glacier wastage was then subtracted from the annual runoff from our calibration gauges in order to avoid double counting this component of glacier runoff in our estimates. Runoff errors were estimated as the standard error of the calibration factor relating PRISM precipitation to runoff in each of the two regions. Glacial wastage and associated errors were calculated using measured glacial wastage for the period mid-1990s to 2001 (ref. 5) from the portion of Alaska/Yukon glaciers that drain to the GOA.

Our glacial discharge estimates (runoff + wastage) are conservative for two reasons: (1) we found that as elevation increased, the PRISM model increasingly underestimated runoff, thus our runoff calibration factors based on watersheds that included some low elevation area probably underestimated runoff from high-elevation glaciers, and (2) recent distributed estimates of glacier wastage from southeast Alaska[28] have shown that current rates of glacial wastage are substantially higher than the estimates from ref. 5 we used. To calculate fluxes of DOC, we used a conservative estimate of the average melt season concentration of DOC in glacial runoff (0.3 mg C l^{-1}) in combination with the modelled glacier water fluxes. The [DOC] estimate we chose was based on our data and previous studies of seasonal DOC dynamics in glacier rivers[12–14]. Unlike northern rivers draining permafrost regions in which DOM is predominantly derived from terrestrial vegetation and soil organic matter[4,26,27], glacier runoff does not exhibit a pronounced spring flush or strong seasonal variability in either the concentration[13,14] or quality[12] of riverine DOM. Thus our DOC flux estimates and associated characterization of the chemical and isotopic composition of glacial DOM are consistent with the current understanding of riverine DOM outputs from glacial ecosystems.

33. Fellman, J. B., Hood, E., D'Amore, D. V. & Boone, R. D. Fluorescence characteristics and biodegradability of dissolved organic matter in forest and wetland soils from coastal temperate watersheds in southeast Alaska. *Biogeochemistry* **88**, 169–184 (2008).
34. Simpson, J. J., Hufford, G. L., Daly, C., Berg, J. S. & Fleming, M. D. Comparing maps of mean monthly surface temperature and precipitation for Alaska and adjacent areas of Canada produced by two different methods. *Arctic* **58**, 137–161 (2005).

LETTERS

Tremor-tide correlations and near-lithostatic pore pressure on the deep San Andreas fault

Amanda M. Thomas[1], Robert M. Nadeau[1] & Roland Bürgmann[1]

Since its initial discovery nearly a decade ago[1], non-volcanic tremor has provided information about a region of the Earth that was previously thought incapable of generating seismic radiation. A thorough explanation of the geologic process responsible for tremor generation has, however, yet to be determined. Owing to their location at the plate interface, temporal correlation with geodetically measured slow-slip events and dominant shear wave energy, tremor observations in southwest Japan have been interpreted as a superposition of many low-frequency earthquakes that represent slip on a fault surface[2,3]. Fluids may also be fundamental to the failure process in subduction zone environments, as teleseismic and tidal modulation of tremor in Cascadia and Japan and high Poisson ratios in both source regions are indicative of pressurized pore fluids[3–7]. Here we identify a robust correlation between extremely small, tidally induced shear stress parallel to the San Andreas fault and non-volcanic tremor activity near Parkfield, California. We suggest that this tremor represents shear failure on a critically stressed fault in the presence of near-lithostatic pore pressure. There are a number of similarities between tremor in subduction zone environments, such as Cascadia and Japan, and tremor on the deep San Andreas transform[3–12], suggesting that the results presented here may also be applicable in other tectonic settings.

Analysis of the response of non-volcanic tremor to small stress oscillations induced in the lithosphere by the Earth's tidal deformation allows the determination of the stress orientations and magnitudes under which tremors preferentially occur, providing additional insight into the frictional processes that control tremor generation. Tidal modulation of tremor was previously established in both Japan and Cascadia[4,6,13,14]. The stresses induced by the ocean tides at these locations are nearly an order of magnitude larger than those near Parkfield[6,14], which primarily arise from the deformation of the solid Earth. We develop our analysis in parallel using a catalogue of 1,777 non-volcanic tremors detected over an eight-year period[11], a regional catalogue of earthquakes within 0.5° of Cholame, the central location of tremor activity, and a catalogue of characteristically repeating micro-earthquakes located along the creeping segment of the San Andreas fault, northwest of Cholame[15] (Fig. 1). We include the repeating earthquake catalogue because it is composed of events that are all located on the creeping San Andreas fault and might thus be expected to experience enhanced tidal triggering compared to regional catalogues[16]. We compute the extensional, shear and volume strains induced by the solid Earth and ocean tides[17,18], assuming a typical tremor source region depth of ~25 km (refs 11, 19). For each catalogue the fault-normal, shear, Coulomb (friction coefficient $\mu = 0.4$), and volumetric stresses and stress rates are computed for the times of each event (see Methods).

To investigate the influence of both the stress magnitude and rate on tremor and earthquake occurrence we divide the catalogue events

into 'quadrants' depending on the sign of the loading condition under which they occur (Fig. 2, inset). If tremor and tides are uncorrelated, the number of events that occur under a particular tidal loading condition will be proportional to the amount of time that particular condition exists. Assuming each tremor event is independent, we use the chi-square statistic to test the null hypothesis that event times are randomly distributed with respect to tidal influence. The results for the tremor and earthquake catalogues are shown in Fig. 2. For the tremor catalogue, the levels of correlation of the normal, shear and Coulomb stresses exceed the 99% significance level while the correlation levels for the other catalogues are statistically insignificant. The lack of correlation in the regional and repeating earthquake catalogues is not surprising given the size of the catalogues and results from previous efforts to establish a significant tidal triggering of earthquakes[20–22] (see Methods).

We further explore the apparent correlation between tremor and tidally induced stresses by comparing tremor times with the loading conditions under which they occur. Figure 3 shows tremor rate distributions with respect to tidally induced shear, normal and Coulomb stress magnitudes and rates at the time of the events. Correlation between tremor occurrence and tidal stressing rate is insignificant for
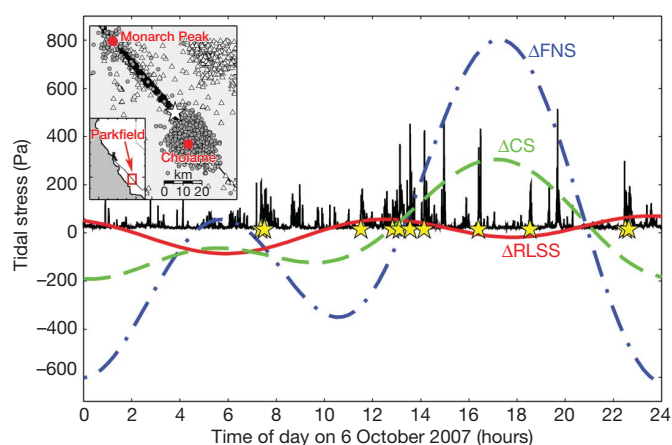

Figure 1 | Example one-day tremor time series with superimposed tidal stresses. Black represents root-mean-square envelope of tremor activity in Cholame. Blue, red and green curves represent the tidally induced fault-normal stress (ΔFNS), right-lateral shear stress (ΔRLSS), and Coulomb stress (ΔCS) for $\mu = 0.4$. Yellow stars mark tremor start times. Some short spikes in the root-mean-square envelope are due to micro-earthquakes. The inset map shows tremor locations as grey circles, locations of regional earthquakes as white triangles, and part of the repeating earthquake catalogue as black squares. Not all repeating events are shown, because they continue further to the northwest along the creeping San Andreas section. The white star indicates the epicentre of the 2004 Parkfield earthquake.

[1]Department of Earth and Planetary Science, Berkeley Seismological Laboratory, University of California-Berkeley, 307 McCone Hall, Berkeley, California 94720–4767, USA.
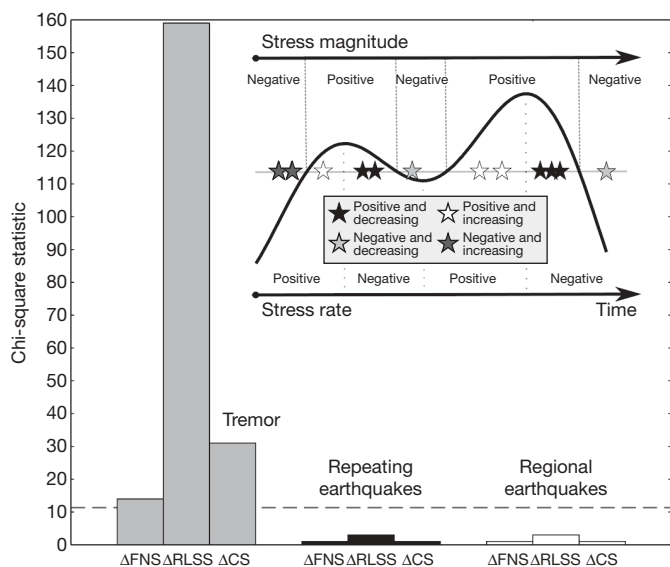
Figure 2 | Results of chi-square significance tests. The chi-square test results correlate the tremor, regional and repeating earthquake catalogues with tidally induced ΔFNS, ΔRLSS, and ΔCS. The dashed line shows the 99% confidence level, where larger values indicate that the hypothesis of random event occurrence can be rejected. The inset shows how tremor start times were separated into four 'quadrants' based on both the magnitude and rate of change of the stresses.

all stress components. Induced right-lateral shear stresses (ΔRLSS) demonstrates the most compelling correlation, with distinct increases in tremor activity that correspond to positive (right-lateral) shear stresses parallel to the San Andreas fault and equally apparent decreases when values are negative. Additionally, the tremor surpluses and deficits become more pronounced as ΔRLSS increases to values of ±150 Pa. Although fault-normal stress changes (ΔFNS) are much larger, they only exhibit a weak correlation at large, positive (tensile) values of >1,000 Pa. Coulomb stresses (ΔCS = μΔFNS + ΔRLSS, μ = 0.4) exhibit less correlation than the shear stress alone.

Assuming a frictional Coulomb failure process is an appropriate model for non-volcanic tremor, the optimal friction coefficient μ is the value that maximizes the number of events that occur during times of encouraged failure stress (Fig. 4, Methods). Tremors show a marked increase for friction values of μ < 0.1 with a peak of over 31% above the long-term average number of events (excess) for μ = 0.02. The percentage excess for both the regional and repeating earthquake catalogues does not exceed 6%. This demonstrates that tidally induced shear stress parallel to the San Andreas fault, although of much smaller magnitude than normal stress changes, has the most robust correlation with non-volcanic tremor near Parkfield. The dominant role of small shear stress perturbations in stimulating the tremor, despite much larger normal stress perturbations, is indicative of a very weak fault zone with low effective normal stress, probably due to near-lithostatic pore pressures at the depth of the tremor source region. As a test of our assumption of San Andreas fault alignment in the stress calculations, we perform the same analysis to determine the percentage excess tremor with respect to a range of vertical fault plane azimuths (Fig. 4, inset). The peak percentage excess occurs at N 44° W, nearly parallel to the local strike of the San Andreas fault (N 45° W). This result serves as an independent piece of evidence to show that tremor represents a mode of fault-parallel shear failure in the lower crust.

If non-volcanic tremor simply represents a different class of earthquakes that still involves failure on a fault surface[13,23], it is unclear why tremor is modulated by tides but regional earthquakes are not. Careful statistical analyses of tidal influence on earthquake populations find the correlation to be extremely weak (~13,000 events are needed to establish a 1% increase)[20-22,24]. Lockner and Beeler[20,21] conducted a series of laboratory tests simulating tidal loading conditions on active faults and found that the absence of strong tidal triggering can be reconciled with rate-state frictional models that involve delayed slip nucleation (for example, by time-dependent failure or static fatigue). If tremor can also be described in a rate-state friction framework then we can derive a quantitative estimate of effective normal stress in the tremor source region from the observed tidal correlations[16]. Dieterich[16] found a simple scaling relationship between earthquake rate fluctuations and the amplitude of the
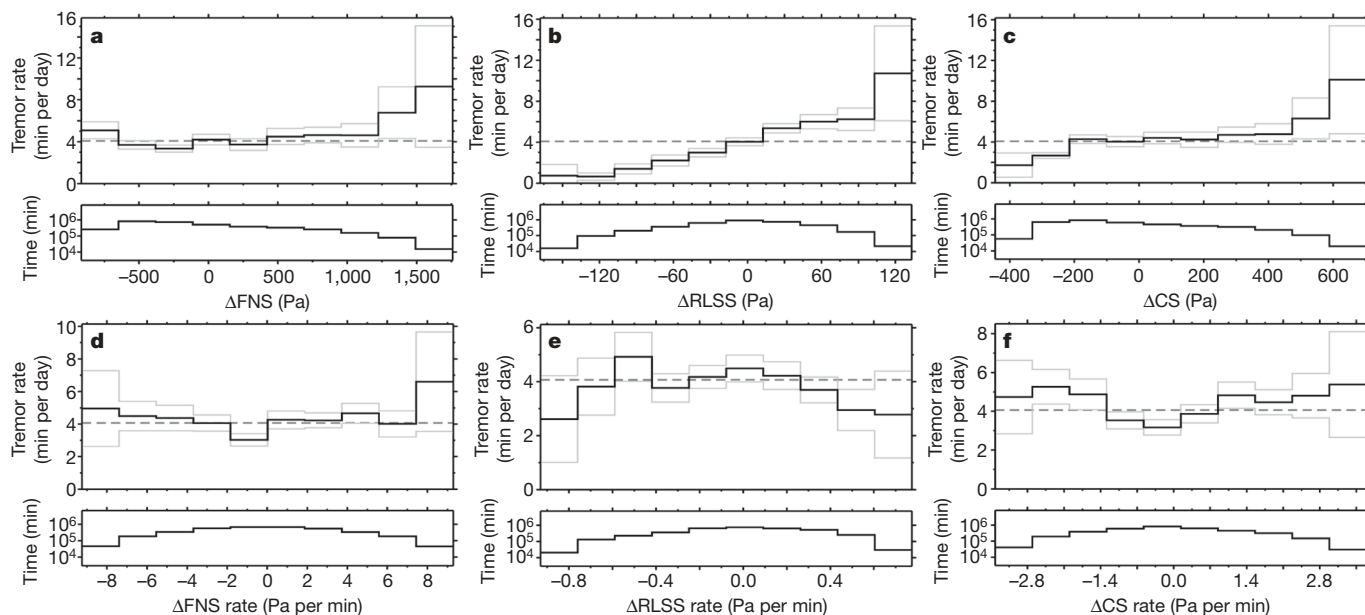

Figure 3 | Tidal stress magnitude and rate distributions. Tidal stress magnitude (panels **a**, **b** and **c**) and rate (panels **d**, **e** and **f**) distributions for the ΔFNS, ΔRLSS and ΔCS (μ = 0.4). Volumetric stresses are not shown because they are strongly coupled with normal stresses. Solid black lines show tremor rates computed by dividing the number of tremors within the respective stress interval by the total amount of time spent in that interval. The time distribution for each bin is shown below each histogram. The average tremor rate (dashed line) corresponds to the rate expected if there were no correlation between tremor and tides. Grey lines are 2σ error bars (Methods).
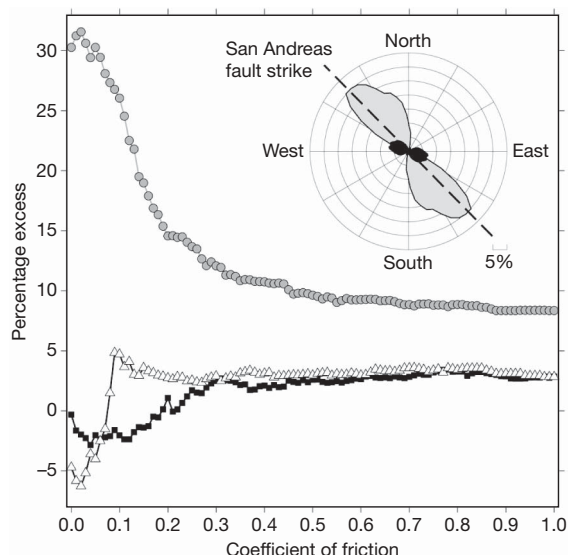
Figure 4 | Percentage of excess events versus friction coefficient.
Percentage of excess events (that is, above long-term average) during times of positive ΔCS versus effective coefficient of friction. Values for the tremor, regional and repeating earthquake catalogues are shown as grey circles, white triangles and black squares respectively. Standard deviations (2σ) were computed using a bootstrap procedure on each friction value for each catalogue. Maximum 2σ errors over all possible friction values are 5.66%, 5.80% and 4.66% for the tremor, regional and repeating earthquake catalogues respectively. The inset displays the positive percentage excess tremor for ΔFNS (black) and ΔRLSS (grey) as a function of the fault azimuth.

periodic shear stress divided by the normal stress:

$$\sigma = 2\tau/(aR_a) \qquad (1)$$

where R_a is the total rate variation divided by the average rate, $R_a = (R_{max} - R_{min})/R_{avg}$, σ is the effective normal stress, τ is the shear stress, and a is the rate constitutive parameter. Using the maximum and minimum rates for $\mu = 0.02$, the value of friction that maximizes the percentage excess (Supplementary Fig. 1), a stressing amplitude of 177 Pa (the maximum value of the tidally induced shear stress) and experimentally derived values of 0.005 to 0.02 for the fault constitutive parameter a (ref. 25), we find effective normal stresses of 0.035 to 0.009 MPa (Methods). These values are orders of magnitude lower than the lithostatic overburden pressure at this depth (~700 MPa), suggesting that effectively lithostatic pore fluids are present in the tremor source region. Additionally, this estimate is consistent with the low friction coefficient reflected in the lack of correlation with larger normal stress fluctuations. Potential sources of fluids at mid-crustal depths beneath the San Andreas fault are less well documented than slab dehydration in subduction zones, but previous studies have suggested the presence of fluids and processes to explain their introduction[26].

The similarities (or differences) in non-volcanic tremor properties with respect to tectonic environment warrant further discussion. Nakata et al.[6] suggest correlation with ΔCS (assuming $\mu = 0.2$) and ΔCS rate in Shikoku, Japan, while Lambert et al.[14] find that peak tremor activity occurs at times of maximum tidal shear stress in the thrust direction and with ΔCS (assuming $\mu = 0.4$) in Cascadia. One possible reason for the different results is that both in Japan (R. Nakata, personal communication, 2009) and Cascadia[14] times of increased shear stress are highly correlated with compressive (clamping) normal stresses of comparable magnitude. This makes it difficult to separate the individual contributions of the normal and shear stresses, especially for the short time periods (a few weeks) considered by these two studies. The results for Parkfield are not subject to this limitation and we find that very small shear stresses, but not the larger normal stress changes, modulate tremor. It is likely that tremor

modulation is affected by frictional strength, the relative magnitude of tectonic and tidal loading rates[20,21], material properties[5], and other factors that vary with respect to tectonic environment. One finding common to all three localities is that tremor appears to be associated with the presence of fluids at near-lithostatic pressures, and given similar observations in variable tectonic environments[3–6], the same mechanism is probably responsible for non-volcanic tremor elsewhere.

We find that tremor occurs preferentially when subjected to tidal shear stresses that promote right-lateral failure along the San Andreas fault, very small stress perturbations from solid-Earth tides are responsible for significant tremor rate increases, and extremely low effective normal stress in the tremor source region is required to explain the apparent tidal triggering. Tremor appears to be shear failure capable of producing seismic radiation, so a brittle failure model seems appropriate. These results present a rheological paradox, because ductile rheologies, which may be expected at depths of more than 20 km and temperatures over 500 °C, generally have no normal-stress dependence. It is possible that very high fluid pressures on discrete fault patches facilitate slip in an otherwise ductile-deforming regime. Coupling these observations with the location of the tremor on deep, roughly planar zones[19], we favour the conjecture that tremor at Cholame represents shear failure within a weak, critically stressed San Andreas fault zone extending to the base of the crust. Finally, given that earthquakes and non-volcanic tremor are both manifestations of slip on a fault surface, further constraining the properties and processes that control tremor generation may facilitate a better understanding of fault mechanics, lithospheric structure and tectonic deformation in the deep crust.

METHODS SUMMARY
Tidal stress computation. Tidally induced strains are computed in the tremor source region using SPOTL[17,18]. Assuming two-dimensional plane strain and linear elasticity, with an elastic modulus of 30 GPa and Poisson ratio of 0.25, strains are then converted to stresses and resolved into fault normal and parallel (shear) directions of the San Andreas fault (N 45° W).
Effective normal stress computation. The stress histogram for friction coefficient $\mu = 0.02$ (the value of the friction coefficient that maximizes the percentage excess from Fig. 4) is shown in Supplementary Fig. 1. From this figure the maximum and minimum tremor rates are 9 min per day and 1 min per day, respectively. Using expression (1) in the text we find that $R_a = 2$, resulting in effective normal stress estimates of 35,400 Pa and 8,850 Pa for a values of 0.005 and 0.02 respectively.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Obara, K. Nonvolcanic deep tremor associated with subduction in southwest Japan. *Science* **296**, 1679–1681 (2002).
2. Shelly, D. R., Beroza, G. C. & Ide, S. Non-volcanic tremor and low-frequency earthquake swarms. *Nature* **446**, 305–307 (2007).
3. Shelly, D. R., Beroza, G. C., Ide, S. & Nakamula, S. Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. *Nature* **442**, 188–191 (2006).
4. Rubinstein, J. L., La Rocca, M., Vidale, J. E., Creager, K. C. & Wech, A. G. Tidal modulation of nonvolcanic tremor. *Science* **319**, 186–189 (2008).
5. Audet, P., Bostock, M. G., Christensen, N. I. & Peacock, S. M. Seismic evidence for overpressured subducted oceanic crust and megathrust fault sealing. *Nature* **457**, 76–78 (2009).
6. Nakata, R., Suda, N. & Tsuruoka, H. Non-volcanic tremor resulting from the combined effect of Earth tides and slow slip events. *Nature Geosci.* **1**, 676–678 (2008).
7. Gomberg, J. et al. Widespread triggering of nonvolcanic tremor in California. *Science* **319**, 173 (2008).
8. Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* **300**, 1942–1943 (2003).
9. Brenguier, F. et al. Postseismic relaxation along the San Andreas fault at Parkfield from continuous seismological observations. *Science* **321**, 1478–1481 (2008).
10. Rubinstein, J. L. et al. Non-volcanic tremor driven by large transient shear stresses. *Nature* **448**, 579–582 (2007).
11. Nadeau, R. M. & Guilhem, A. Nonvolcanic tremor evolution and the San Simeon and Parkfield, California, earthquakes. *Science* **325**, 191–193 (2009).
12. Nadeau, R. M. & Dolenc, D. Nonvolcanic tremors deep beneath the San Andreas fault. *Science* **307**, 389 (2005).

13. Shelly, D. R., Beroza, G. C. & Ide, S. Complex evolution of transient slip derived from precise tremor locations in western Shikoku, Japan. *Geochem. Geophys. Geosyst.* **8**, doi:10.1029/2007gc001640 (2007).

14. Lambert, A., Kao, H., Rogers, G. & Courtier, N. Correlation of tremor activity with tidal stress in the northern Cascadia subduction zone. *J. Geophys. Res. Solid Earth* **114**, doi:10.1029/2008JB006038 (2009).

15. Nadeau, R. M. & McEvilly, T. V. Periodic pulsing of characteristic microearthquakes on the San Andreas fault. *Science* **303**, 220–222 (2004).

16. Dieterich, J. H. Nucleation and triggering of earthquake slip—effect of periodic stresses. *Tectonophysics* **144**, 127–139 (1987).

17. Agnew, D. *SPOTL: Some Programs for Ocean Tide Loading* ⟨http://igppweb.ucsd.edu/~agnew/spotlmain.html⟩ (1996).

18. Agnew, D. C. NLOADF: A program for computing ocean-tide loading. *J. Geophys. Res. Solid Earth* **102**, 5109–5110 (1997).

19. Shelly, D. R. *et al.* Precise location of San Andreas fault tremors near Cholame, California using seismometer clusters: slip on the deep extension of the fault? *Geophys. Res. Lett.* **36**, doi:10.1029/2008gl036367 (2009).

20. Beeler, N. M. & Lockner, D. A. Why earthquakes correlate weakly with the solid Earth tides: effects of periodic stress on the rate and probability of earthquake occurrence. *J. Geophys. Res. Solid Earth* **108**, doi:10.1029/2001jb001518 (2003).

21. Lockner, D. A. & Beeler, N. M. Premonitory slip and tidal triggering of earthquakes. *J. Geophys. Res. Solid Earth* **104**, 20133–20151 (1999).

22. Vidale, J. E., Agnew, D. C., Johnston, M. J. S. & Oppenheimer, D. H. Absence of earthquake correlation with Earth tides: an indication of high preseismic fault stress rate. *J. Geophys. Res. Solid Earth* **103**, 24567–24572 (1998).

23. Ide, S., Shelly, D. R. & Beroza, G. C. Mechanism of deep low frequency earthquakes: further evidence that deep non-volcanic tremor is generated by shear slip on the plate interface. *Geophys. Res. Lett.* L03308 doi:10.1029/2006gl028890 (2007).

24. Cochran, E. S., Vidale, J. E. & Tanaka, S. Earth tides can trigger shallow thrust fault earthquakes. *Science* **306**, 1164–1166 (2004).

25. Blanpied, M. L., Lockner, D. A. & Byerlee, J. D. Frictional slip of granite at hydrothermal conditions. *J. Geophys. Res. Solid Earth* **100**, 13045–13064 (1995).

26. Becken, M. *et al.* A deep crustal fluid channel into the San Andreas Fault System near Parkfield, California. *Geophys. J. Int.* **173**, 718–732 (2008).

Author Contributions A.M.T. performed the analysis. A.M.T. and R.B. wrote the paper. R.M.N. developed the tremor and repeating earthquake catalogues and performed analysis related to Figure 3. All authors contributed to the interpretation and final manuscript preparation.

Author Information The authors declare no competing financial interests. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M.T. (amthomas@berkeley.edu).

METHODS

Event catalogues. The tremor catalogue includes a total of 1,777 events between July 2001 and May 2008 using the detection methodology described in ref. 11. A small portion (5–10%) of the non-volcanic tremor occur farther north beneath Monarch Peak, but the tidally induced stresses do not vary in a significant way over such small distances, so tidal-stress time series were computed for Cholame only. The repeating earthquake catalogue contains 2,594 events between 1984 and 1999 on the 175-km creeping section of the San Andreas fault[15]. The tidal stresses for this catalogue were computed at the centre of the creeping section at 5 km depth. Finally, the regional earthquake catalogue consists of all earthquakes from the Advanced National Seismic System catalogue, for the same time period as the tremor catalogue, that are within 0.5° from the centre of the Cholame tremor. The same tidal time series were used for the tremor and regional earthquakes and were computed at 35.666° N, 120.2854° W.

Tidal stress calculation. The tidally induced stresses in the lithosphere were computed using the SPOTL code developed by D. Agnew[17,18]. SPOTL uses Green's functions to compute azimuthal and vertical strains that can subsequently be converted to stress. A more thorough description of the mathematical treatment of tides and code specifics can be found on the SPOTL webpage (http://www.igpp.ucsd.edu/~agnew/spotlmain.html). The results in this analysis include contributions from both the body tides, which arise due to deformation of the solid Earth, and the ocean load tides. The computation includes the semi-diurnal M2, N2, S2 and K2, and diurnal K1, O1, P1 and Q1 tidal constituents and we used the TOPEX/Poseidon global ocean tide model included in the SPOTL package to compute the load tides. Tidal model predictions were compared with strain-meter records in Piñon Flat using the perturbation matrix of ref. 27, as well as previous studies to verify the calculation[22,27].

Statistical methods. Pearson's chi-square test is designed to test the similarity between two frequency distributions[28]. The measure of similarity is known as the chi-squared statistic and is defined as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where n is the total number of potential outcomes, O_i is the observed frequency for a particular outcome i, and E_i is the expected frequency for a particular outcome[28]. In the case of non-volcanic tremor, the total number of tremors are each assigned to one of four possible categories or 'quadrants' (positive and decreasing, negative and decreasing, and so on) depending on the behaviour of the tidally induced stress under consideration at the time the tremor occurred.

The expected number is computed by taking the entire eight-year tidal time series and determining what fraction of time was spent in each group, then multiplying by the total number of tremors in the catalogue. In this way we account for the fact that the time distribution is not equivalent between quadrants. The grey dashed line in Fig. 2 is the critical value (99% significance level) for an upper one-sided test with three degrees of freedom.

Each tremor in Fig. 3 is weighted by its duration so, for example, one 15-min tremor is equivalent to three 5-min tremors. This is in contrast to Figs 2 and 4, where all tremors are weighted equally. We chose this approach because the tremor rates can then be used to compute an effective normal stress in the tremor source region.

Figure 4 again weights all tremors equally and is somewhat analogous to the chi-square tests mentioned earlier. The percentage excess is defined as $(O - E)/E$, using the same observed and expected values (O and E) as in the chi-square test, but in the case of Fig. 4 we consider only the positive magnitudes ($n = 1$). The percentage excess is roughly related to the chi-square values, the main difference being that the chi-square is a measure of the variance of a population whereas the percentage excess is related to the mean. Two populations of different size with the same percentage excess (10 coin flips with 6 heads and 1,000 coin flips with 600 heads, for example) will have greatly different chi-square statistics. The larger population will have the highest chi-square value because it is much more statistically improbable.

Uncertainty estimates from Fig. 3. Two-standard-deviation error estimates are determined using a bootstrap method. The method randomly selects an individual tremor from the original tremor catalogue of 1,777 tremors. This process is repeated 1,777 times to yield a randomly sampled catalogue (with replacement) of 1,777 events. The above process is repeated 50 times, giving 50 randomly sampled versions of the original catalogue. The tremor rate distributions as a function of the tidally induced stresses are then calculated for each of the 50 randomly sampled catalogues. For each stress bin, the standard deviation of tremor rates of the 50 catalogue values are then computed and multiplied by 2, giving the 2σ rate uncertainties for each bin that are shown by the grey bars in Fig. 3. Tremor rate errors arise from variability in tremor detection sensitivity. More details on the detection methodology can be found in ref. 11.

27. Hart, R. H. G. *et al.* Tidal calibration of borehole strainmeters: removing the effects of small-scale inhomogeneity. *J. Geophys. Res.* 101, 25553–25571 (1996).

28. Devore, J. L. *Probability and Statistics: for Engineering and the Sciences* 6th edn, Ch. 14 (Brooks/Cole-Thomson Learning, 2004).

LETTERS

The velocity of climate change

Scott R. Loarie[1], Philip B. Duffy[1,2], Healy Hamilton[3], Gregory P. Asner[1], Christopher B. Field[1] & David D. Ackerly[4]

The ranges of plants and animals are moving in response to recent changes in climate[1]. As temperatures rise, ecosystems with 'nowhere to go', such as mountains, are considered to be more threatened[2,3]. However, species survival may depend as much on keeping pace with moving climates as the climate's ultimate persistence[4,5]. Here we present a new index of the velocity of temperature change (km yr^{-1}), derived from spatial gradients (°C km^{-1}) and multimodel ensemble forecasts of rates of temperature increase (°C yr^{-1}) in the twenty-first century. This index represents the instantaneous local velocity along Earth's surface needed to maintain constant temperatures, and has a global mean of 0.42 km yr^{-1} (A1B emission scenario). Owing to topographic effects, the velocity of temperature change is lowest in mountainous biomes such as tropical and subtropical coniferous forests (0.08 km yr^{-1}), temperate coniferous forest, and montane grasslands. Velocities are highest in flooded grasslands (1.26 km yr^{-1}), mangroves and deserts. High velocities suggest that the climates of only 8% of global protected areas have residence times exceeding 100 years. Small protected areas exacerbate the problem in Mediterranean-type and temperate coniferous forest biomes. Large protected areas may mitigate the problem in desert biomes. These results indicate management strategies for minimizing biodiversity loss from climate change. Montane landscapes may effectively shelter many species into the next century. Elsewhere, reduced emissions, a much expanded network of protected areas[6], or efforts to increase species movement may be necessary[7].

As climate changes in this century, the current distribution of climatic conditions will be rearranged on the globe; some climates will disappear entirely, and new (no-analogue) climates are expected in wide regions[8]. For species to survive, the persistence of suitable climates is not sufficient. Species must also keep pace with climates as they move[9]. To summarize the speed at which climate is changing, we compute the instantaneous horizontal velocity of temperature change (Fig. 1e) derived from the ratio of temporal (Fig. 1d) and spatial (Fig. 1c) gradients of mean annual near-surface temperature (Fig. 1b) (°C yr^{-1}/°C km^{-1} = km yr^{-1}). As exemplified by California, the spatial gradient of temperature change is greatest on mountain slopes (Fig. 1c): modest displacements in space, moving up or down slope, result in a large change in temperature. As a result, relatively small velocities (Fig. 1e) are required to keep pace with the rate of temperature change (assuming that the climate persists and does not disappear). In contrast, high velocities are required in flatter areas such as California's Central Valley, where large geographic displacements are required to change temperature appreciably.

The magnitude of these results depends on the emissions scenario (Fig. 2a) and also on the period of time over which the temporal gradient is calculated (Fig. 2b). However, velocity patterns of global temperature change are similar across scenarios, with the highest velocities in flat landscapes at higher latitudes (Fig. 2c). Using temperature change calculated from 2000–2100 under the intermediate

A1B emissions scenario, the geometric mean velocity was 0.42 km yr^{-1} (0.11–1.46). (Throughout, we summarize uncertainty in the mean by listing upper and lower, ± 1 s.d., estimates in parenthesis.) See Supplementary Fig. 17 for other emissions scenarios. We summarize velocity for biomes of the globe and rank them by increasing mean velocity (Fig. 3). Doing so shows that mountainous biomes require the slowest velocities to keep pace with climate change. In contrast, flatter biomes such as flooded grasslands, mangroves and deserts require much greater velocities. Overall, there is a strong correlation between topographic slope and velocity from temperature change (correlation coefficient = −0.92; see Supplementary Fig. 18).

Land-cover change results in formidable barriers to species movement[10]. Thus, keeping pace with climate change is more feasible in protected areas where landscapes may be less fragmented[11]. The sizes
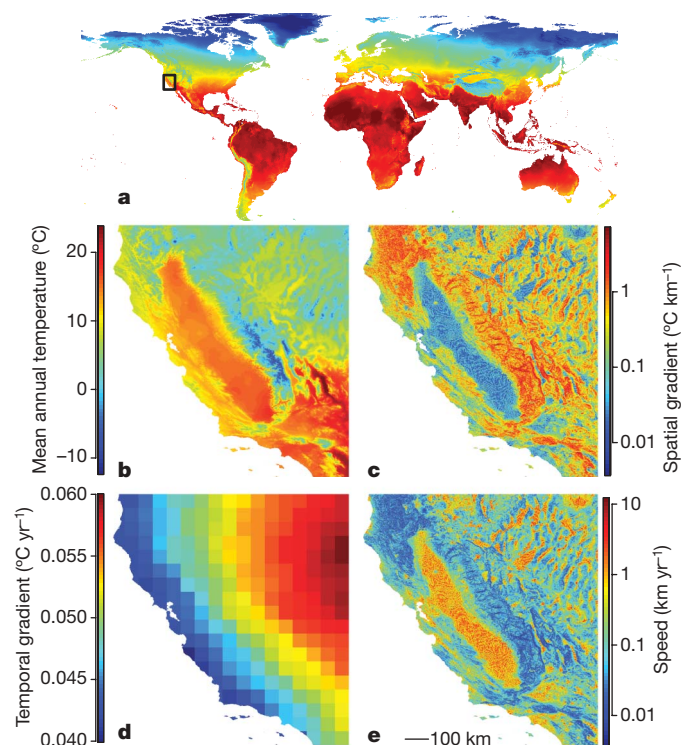

Figure 1 | Changing temperature in California. a, Current (1950–2000) mean annual temperature at 800 m resolution. The black rectangle indicates the Central California inset in **b. c,** The spatial gradient of temperature change using a 9 pixel kernel. **d,** The temporal gradient of climate change from 2000–2099 from 0.5 °C 16 general circulation model (GCM) ensemble projection with A1B emissions. **e,** The velocity of climate change determined from the quotient of **d** and **c**.

[1]Carnegie Institution for Science, Department of Global Ecology, Stanford, California 94305, USA. [2]Climate Central, Inc., Palo Alto, California 94301, USA. [3]Center for Applied Biodiversity Informatics, California Academy of Sciences, San Francisco, California 94118, USA. [4]Department of Integrative Biology, University of California, Berkeley, California 94720, USA.
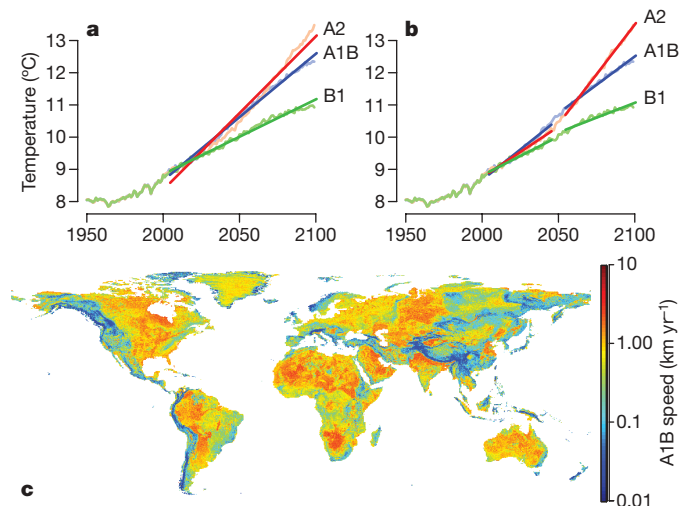
Figure 2 | The velocity of temperature change globally. a, Temporal gradients calculated from 2000–2100 across three emissions scenarios (A2, A1B and B1). **b**, Temporal gradients calculated from 2000–2050 and 2050–2100 across three emissions scenarios. Trends plotted here are the average of the global land surface. **c**, A global map of climate velocity calculated using the 2050–2100 Special Report on Emissions Scenarios (SRES) A1B emissions scenario temporal gradient.

areas are circular and disconnected, this index can be interpreted as the time for current climate to cross a protected area. Such residence times exceed 100 years for only 8.02% (2.67–16.49) of protected areas. Figure 4 summarizes these data by biome, ordered by decreasing residence time. The order of residence times is generally the inverse of velocities, across biomes. For example, the three biomes with the slowest velocities have among the four longest residence times. There are also notable differences. For example, the limited size of protected areas in Mediterranean-type, temperate broadleaf and coniferous forest biomes decreases the residence time in these biomes despite relatively low velocities. The rank of these biomes increased from 4, 7 and 2 to 11, 13 and 8 when arranged by decreasing mean residence time (Fig. 4) as opposed to increasing mean velocity (Fig. 3). In contrast, larger protected areas in other biomes decreased their rank despite high velocities of temperature change. For example, the deserts decreased in rank from 12 to 6.

To guide the interpretation of these results, we make three clarifications. First, climate change involves complex interactions among temperature, precipitation and seasonal and historic variability. We focus on mean annual temperature for several reasons. Mean annual temperature is a useful summary of both historic and projected climate change. The direction and magnitude of temperature change is much less uncertain than precipitation change[12]. Growing numbers of examples document the latitudes and elevations of species distributions responding as expected to changing temperatures[13–16]. We repeated all analyses with precipitation (see Supplementary Fig. 19). Interestingly, precipitation spatial gradients are also greatest in mountainous areas due to the influence of rain shadows and orographic effects. As a result the mean velocity, 0.22 km yr^{-1} (0.08–1.90), and overall patterns are similar to those derived by temperature.

Second, there is uncertainty in both the spatial and temporal gradients of climate change. We estimated uncertainty in both of these contributing factors and propagated them through to lower and upper estimates of velocity and residence time (see Supplementary

of protected areas vary greatly across biomes (see Supplementary Fig. 20). To explore the interaction between protected area sizes and velocities required to keep pace with climate change, we calculated residence times, defined as the diameter of each protected area divided by velocity (km/km yr^{-1} = yr). Assuming that protected
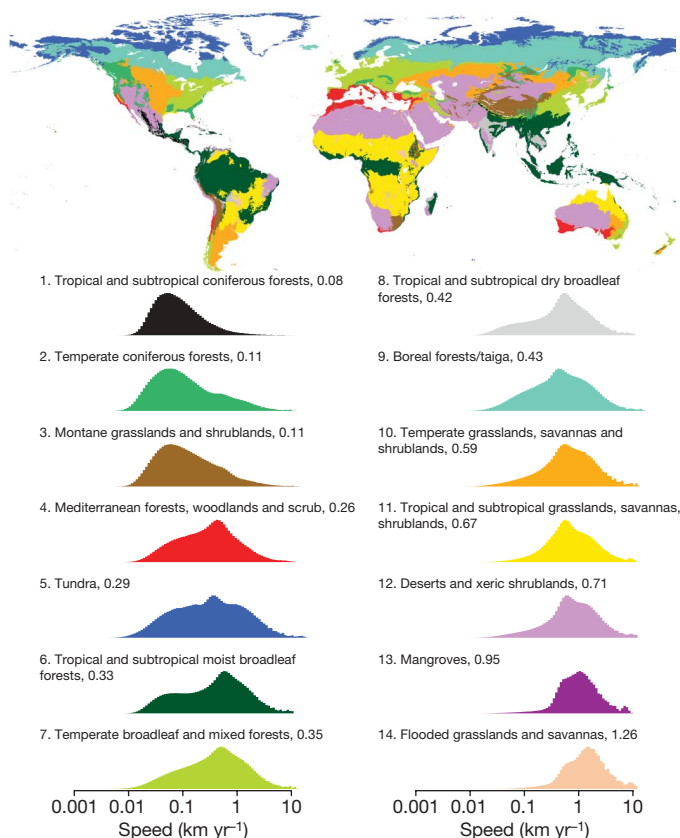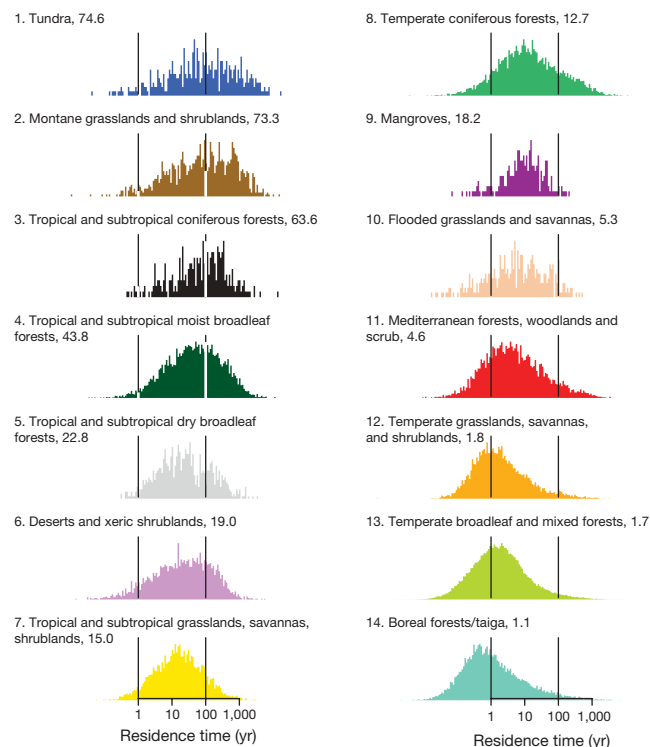

Figure 3 | The velocity of temperature change by biome. A map of biomes and histograms of the speed of temperature change within each biome. Histograms are ordered by increasing velocity according to their geometric means.

Figure 4 | Climate residence time (yr) in protected areas. Histograms represent the ratio of protected area diameter (km) to climate velocity (km h^{-1}), and are ordered by decreasing mean residence time across biomes. The vertical bar indicates 1 and 100 years.

Materials). Furthermore, we note the instantaneous velocity is sensitive to the grain of the analysis and the size of the kernels used to compute spatial and temporal gradients[17]. Throughout this study, we use a ~1 km spatial grain size, the finest available for global mean annual temperature, and a 9 grid cell spatial kernel, standard for creating gridded slopes from digital elevation models. We chose the fine spatial grain because both yearly dispersal capabilities for many species and large temperature changes often occur on the order of a few kilometres. In some cases, even finer resolution responses on the scale of metres may be sufficient to offset climate change as populations move from south to north facing slopes within a habitat[18].

Third, our index estimates the velocities and residence times of climates, not species. We make no assumptions about the tolerances of individual species. Implications for individual species depend on the breadth of tolerance[15], and our velocities apply to range edges, ecotones[19] and other features that reflect climate isoclines. For species with small tolerances, the velocity estimates closely approximate migration speeds required to potentially avoid extinction. For species with large tolerances, the residence times are underestimates. We also note that species do not move at constant rates[20] and, in some instances, the velocity of movements downhill may differ from those uphill; similarly leading-edge expansion and trailing-edge contraction will be different, reflecting the contrasting mechanisms. As a result of these caveats, we interpret these velocities as a relative index of the speeds required to keep pace with climate change rather than a calibrated index of migration rates.

It is, however, interesting to compare the magnitude and spatial patterns of the velocities with migration rate studies. A previous study[21] calculated the minimum distances between modelled current and future biomes from two Global Vegetation Models (GVMs) and interpreted them as necessary speeds for species migration. The study used similar emissions and temporal scales, a much coarser spatial grain (0.5 °C), and projections from an older generation of global climate models than those explored here. Accounting for uncertainty, the probability density function of our velocities are the same as those found previously[21]. For example, we calculated that 28.8% (0.5–66.9) of the globe had migration rates greater than $1 \, \mathrm{km} \, \mathrm{yr}^{-1}$, compared with 17.4% and 21.1% from the two GVMs used previously[21]. These similarities are interesting given the many differences in the approaches used to explore migration rates (see Supplementary Material for further comparisons).

The rate of northward tree migration during the Holocene is estimated at about $1 \, \mathrm{km} \, \mathrm{yr}^{-1}$ after the last glacial maximum in Europe and North America[9]. The apparent paradox of such a fast migration rate relative to the limitations on plant dispersal[22] is possible by rare long-distance migration events[23] or high latitude refugia reseeding the landscape[24]. The latter means that post-glacial re-colonization velocities may have been as much as an order of magnitude slower than previously thought ($\sim 0.1 \, \mathrm{km} \, \mathrm{yr}^{-1}$).

We project that large areas of the globe (28.8%) will require velocities faster than the more optimistic plant migration estimates from a landscape before anthropogenic fragmentation. Velocities on montane landscapes, in contrast, may fall within historic rates. The ability of complex topography to provide a spatial buffer for climate change has been recognized qualitatively[25] and evaluated over small geographic areas[26], but was muted in previous, coarser scale global analyses[21]. Considering these factors adds an important dimension to management strategies for addressing climate change, highlighting the greater vulnerability of large, extensive areas such as the lowland tropics and desert regions. In landscapes where small velocities are required, moderate-sized protected areas may be able to contain moving climates and ecosystems. Elsewhere, further steps must be taken. These include slowing the temporal gradient of climate change through reduced emissions, increasing the ability of plants and animals to disperse through managed relocation[7], or increasing the size of protected areas through habitat corridors and new reserves[6].

METHODS SUMMARY

In brief, for current climate, we used the 30-arcsec WorldClim Version 1.4 Annual Mean Temperature and Total Annual Precipitation bioclimatic variable[27]. Spatial gradients were calculated from a 3×3 grid cell neighbourhood using the average maximum technique[28] modified to accommodate different cell widths at different latitudes. Future temperature projections for each emissions scenario were averages of 16 global climate models statistically downscaled to 0.5 °C[29]. We computed temporal gradients as the slope of a linear model fit through each year of the time period of interest. Velocity is the ratio of the temporal gradient to the spatial gradient, and log transformed for visualization due to a highly skewed distribution. We report geometric means of the velocity to accommodate this skew. We compiled biomes from the World Wildlife Fund (WWF) Terrestrial Ecoregions[30] and protected areas from the World Database on Protected Areas (WDPA) Annual Release 2009 (web download version), February 2009. For each protected area centroid we sampled the velocity and biome to compare with estimates of reserve diameter. We discuss precipitation and uncertainty propagation in the Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Parmesan, C. & Yohe, G. A globally coherent fingerprint of climate change impacts across natural systems. *Nature* **421**, 37–42 (2003).
2. Nogués-Bravo, D., Araujo, M. B., Errea, M. P. & Martinez-Rica, J. P. Exposure of global mountain systems to climate warming during the 21st Century. *Glob. Environ. Change* **17**, 420–428 (2007).
3. Colwell, R. K., Brehm, G., Cardelus, C. L., Gilman, A. C. & Longino, J. T. Global warming, elevational range shifts, and lowland biotic attrition in the wet tropics. *Science* **322**, 258–261 (2008).
4. Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. & Prentice, I. C. Climate change threats to plant diversity in Europe. *Proc. Natl Acad. Sci. USA* **102**, 8245–8250 (2005).
5. Loarie, S. R. et al. Climate change and the future of California's endemic flora. *PLoS One* **3**, e2502 (2008).
6. Hannah, L. Protected areas and climate change. *Ann. NY Acad. Sci.* **1134**, 201–212 (2008).
7. Hoegh-Guldberg, O. et al. Assisted colonization and rapid climate change. *Science* **321**, 345–346 (2008).
8. Williams, J. W., Jackson, S. T. & Kutzbach, J. E. Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl Acad. Sci. USA* **104**, 5738–5742 (2007).
9. Pearson, R. G. Climate change and the migration capacity of species. *Trends Ecol. Evol.* **21**, 111–113 (2006).
10. Damschen, E. I., Haddad, N. M., Orrock, J. L., Tewksbury, J. J. & Levey, D. J. Corridors increase plant species richness at large scales. *Science* **313**, 1284–1286 (2006).
11. Joppa, L. N., Loarie, S. R. & Pimm, S. L. On the protection of "protected areas". *Proc. Natl Acad. Sci. USA* **105**, 6673–6678 (2008).
12. Murphy, J. M. et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772 (2004).
13. Moritz, C. et al. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. *Science* **322**, 261–264 (2008).
14. Kelly, A. E. & Goulden, M. L. Rapid shifts in plant distribution with recent climate change. *Proc. Natl Acad. Sci. USA* **105**, 11823–11826 (2008).
15. Parmesan, C. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669 (2006).
16. Lenoir, J., Gegout, J. C., Marquet, P. A., De Ruffray, P. & Brisse, H. A significant upward shift in plant species optimum elevation during the 20th century. *Science* **320**, 1768–1771 (2008).
17. Trivedi, M. R., Berry, P. M., Morecroft, M. D. & Dawson, T. P. Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. *Glob. Change Biol.* **14**, 1089–1103 (2008).
18. Weiss, S. B., Murphy, D. D. & White, R. R. Sun, slope, and butterflies: topographic determinants of habitat quality for *Euphydryas editha*. *Ecology* **69**, 1486–1496 (1988).
19. Beckage, B. et al. A rapid upward shift of a forest ecotone during 40 years of warming in the Green Mountains of Vermont. *Proc. Natl Acad. Sci. USA* **105**, 4197–4202 (2008).
20. Van Houtan, K. S., Pimm, S. L., Halley, J. M., Bierregaard, R. O. & Lovejoy, T. E. Dispersal of Amazonian birds in continuous and fragmented forest. *Ecol. Lett.* **10**, 219–229 (2007).
21. Malcolm, J. R., Markham, A., Neilson, R. P. & Garaci, M. Estimated migration rates under scenarios of global climate change. *J. Biogeogr.* **29**, 835–849 (2002).
22. Clark, J. S. et al. Reid's paradox of rapid plant migration. *Bioscience* **48**, 13–24 (1998).
23. Clark, J. S. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *Am. Nat.* **152**, 204–224 (1998).

24. McLachlan, J. S., Clark, J. S. & Manos, P. S. Molecular indicators of tree migration capacity under rapid climate change. *Ecology* **86**, 2088–2098 (2005).
25. Peterson, D. L., Schreiner, E. G. & Buckingham, N. M. Gradients, vegetation and climate: spatial and temporal dynamics in the Olympic Mountains, USA. *Glob. Ecol. Biogeogr. Lett.* **6**, 7–17 (1997).
26. Peterson, A. T. Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Glob. Change Biol.* **9**, 647–655 (2003).
27. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
28. Burrough, P. A. & McDonnell, R. A. *Principles of GIS* 190 (Oxford Univ. Press, 1998).
29. Maurer, E. P., Adam, J. C. & Wood, A. W. Climate model based consensus on the hydrologic impacts of climate change to the Rio Lempa basin of Central America. *Hydrol. Earth Syst. Sci.* **13**, 183–194 (2009).
30. Olson, D. M. *et al.* Terrestrial ecoregions of the world: a new map of life on earth. *Bioscience* **51**, 933–938 (2001).

METHODS

Temporal gradients. We performed all analyses using both mean annual temperature (°C) and total annual precipitation (mm). Future temperature and precipitation projections for each emissions scenario were averages of 16 GCMs statistically downscaled to 0.5 °C[29]. Exactly one ensemble member was used from each GCM (Supplementary Table 1).

To compute temporal gradients, we first summarized the 16 GCM time series for each pixel (for a given emission scenario) into mean, upper and lower time series by calculating the mean and the mean ± 1 s.d. of the 16 projections at each year in the time series (Supplementary Fig. 1, dashed black lines and grey bounds). We then calculated the mean slope and intercept from this mean time series using linear regression (Supplementary Fig. 1, solid black line). Next, we approximated upper and lower slopes as the difference between the height of the mean regression line at 2000 and the height of the upper and lower time series at 2100 divided by 101 years (Supplementary Fig. 1, red and blue lines).

We summarize uncertainty by mapping these upper and lower gradients globally for temperature (Supplementary Fig. 2) and precipitation (Supplementary Fig. 3). We plot their histograms (Supplementary Figs 4 and 5) grouped by biome. Biomes were compiled from the WWF Terrestrial Ecoregions[30]. Interestingly, because in many places the mean change in precipitation is close to zero but the uncertainty is large, in many cases both the upper and lower uncertainty bounds on the precipitation temporal gradients are larger than the mean.

There are several different approaches that could be used to estimate temporal gradients. One alternative approach would be to first calculate separate slopes for each pixel from each of the 16 separate GCM time series and subsequently calculate the mean and s.d. from these 16 slopes. To explore whether selecting this alternative approach would influence estimates of temporal gradients and uncertainty, we compared both approaches for a single arbitrary pixel in California (Supplementary Fig. 6). In the graph illustrating the alternative approach (Supplementary Fig. 6, right column), we arbitrarily fix the height of the upper and lower lines in the year 2000 to the height of the mean line. For the pixel in Supplementary Fig. 6, under the SRES A1B emission scenario simulations from 2000–2100, the two approaches yield similar results with a temperature temporal gradient of 0.03 °C yr^{-1} (0.02–0.04) for both approaches, and a precipitation temporal gradient of 0.14 (1.66–2.69) for the approach used here and 0.14 mm yr^{-1} (1.20–1.48) for the alternative approach.

Next, we explored whether the global land surface distribution of temporal gradients was influenced by the approach used. Supplementary Fig. 7 shows global land surface histograms of the mean (black), lower (blue) and upper (red) temporal gradient estimates for temperature and precipitation calculated from each of the two approaches under the SRES A1B emission scenario simulations from 2000–2100. The global results are very similar regardless of the approach used with average temperature temporal gradients of 0.04 °C yr^{-1} (0.03–0.05) for either approach, and average precipitation temporal gradients of 0.59 mm yr^{-1} (0.62–1.72) for the approach used here and 0.55 mm yr^{-1} (0.54–1.06) for the alternative approach. We chose the approach used here because we found it easier to visualize.

Spatial gradients. We performed all analyses globally at 30 arcsec (~0.8 km). We used the 30-arcsec WorldClim Version 1.4 (ref. 27) Annual Mean Temperature and Annual Total Precipitation bioclimatic variables[27] compiled from 1950–2000 to approximate current climate. To decrease the incidence of flat spatial gradients that cause infinite speeds, we added uniformly distributed random noise from −0.05 to 0.05 °C, or from −0.5 to 0.5 mm, to each pixel for temperature and precipitation, respectively.

From the current climate maps, we calculated spatial gradients from a 3×3 grid cell neighbourhood using the average maximum technique[28] modified to accommodate different cell widths at different latitudes. To convert cell height in latitudinal degrees to km, we used 111.325 km degree^{-1}. To convert cell width in longitudinal degrees to km we calculated $\cos\left(\frac{\pi}{180}y\right)111.325$, in which y is the latitude of the pixel in degrees.

Uncertainty in WorldClim temperature and precipitation varies spatially but is around 0.1 °C and 10 mm[3]. Using the analogous PRISM dataset[31] over the western United States as a reference (Supplementary Fig. 8), we determined the difference between WorldClim and PRISM to be similar to the reported uncertainty. Propagating uncertainty in WorldClim through to the spatial gradient would overestimate the uncertainty because the spatial gradient is derived from relative climate and not absolute magnitudes (Supplementary Fig. 9). Accordingly, we estimate uncertainty in the spatial gradient directly. First, we note that standard deviation of spatial gradients within WorldClim varies by topographic Digital Elevation Model (DEM) slope. To accommodate this, we calculate empirical standard deviations binned by topographic slope and fit a smoothed spline through them (Supplementary Fig. 10, green lines). We divided

the log of topographic slope (m km^{-1}) into bins of size 0.1, in which each bin, j, contained I_j pixels. Because the topographic slope is an input into the WorldClim model, these standard deviations may underestimate actual uncertainty in the spatial gradient. Instead, we use PRISM as a reference and calculate the s.d. for each bin as

$$\sigma_j = \sqrt{\frac{1}{I_j - 1}\sum_{i=1}^{I_j}(W_{j,i} - P_{j,i})^2}$$

between the WorldClim spatial gradient, $W_{j,i}$, and the paired PRISM, $P_{j,i}$, spatial gradient at each pixel, i, in each bin, j. We use the smoothed spline fit through these estimates as the one standard deviation upper and lower uncertainty bounds on the spatial gradient (Supplementary Fig. 10, red and blue lines). We map this uncertainty globally for temperature (Supplementary Fig. 11) and precipitation (Supplementary Fig. 12). We note that although we apply uncertainty derived from the western United States across the globe, weather station density is lower in less developed countries. This means that uncertainty is probably higher in such countries especially if they are mountainous. The development of higher accuracy global weather surfaces with known uncertainties is a priority for better understanding the impact of climate change on ecosystems.

Velocity calculations. To combine the 0.5 °C (~50 km) temporal gradient maps with the 1 km spatial gradient maps, we resampled the temporal gradient maps to 1 km. We calculated speed as the ratio of the temporal gradient to the spatial gradient. We only computed speed in cells with data on both spatial and temporal gradients. This excluded certain areas along the coast that were not included in the 0.5 °C maps.

To estimate uncertainty in the velocity calculations, we propagated the upper and lower uncertainty in the temporal and spatial gradients. Again, we map this uncertainty globally for temperature (Supplementary Fig. 13) and precipitation (Supplementary Fig. 14), and plot their histograms grouped by biome (Supplementary Figs 15 and 16). For visualization, we log transformed the histograms due to a highly skewed distribution. In the manuscript text, we report geometric means of the speed index to accommodate this skew. Infinite speeds (resulting from a spatial gradient of zero) were set to the global finite 99.99% quantile value for the purposes of plotting.

We recognize that combining gradients with different measurement scales and units (spatial and temporal) into a single index raises several uncertainties and possible issues. First, as we state in the manuscript, the kernels used to produce gradients are arbitrary and cannot be made directly equivalent. We use temporal gradients from 2000–2100 but explore gradients from 2000–2050 and 2050–2100. We use the finest spatial gradient possible globally—a grid of 9 grid cells—based on our argument of the potential important role of microclimates. Second, extrapolating these velocities over time or space must be done with caution, as doing so assumes that the gradients extrapolate linearly. In reality, topography can change suddenly as plains give way to mountains and vice versa. Likewise, depending on emission scenarios and climate feedbacks, climate change will probably not proceed linearly over the next century. We have purposely left differences in emission scenarios out of our uncertainty analysis (as opposed to differences in global climate models) because future climate policy will give society varying degrees of control over future emissions.

We repeated the analyses of the velocity of temperature change (km yr^{-1}) using the 2050–2100 SRES B1 and SRES A2 emissions scenario temporal gradient (Supplementary Fig. 17). Increasing emissions from the milder B1 to the more severe A2 scenario increases mean velocities from 0.28 to 0.57 km yr^{-1}.

To explore the relationship between topographic slope and the velocity of climate change, we plotted the speed (km yr^{-1}) of temperature and precipitation change by topographic slope (m km^{-1}) using the WorldClim DEM[27] and computed the correlation (Supplementary Fig. 18). Red points are speeds of temperature change for each pixel (correlation coefficient = −0.92). Black points are speeds of precipitation change (correlation coefficient = −0.46). The correlation is stronger for temperature.

We map areas where the speed from precipitation change is greater than the speed from temperature change as well as the average magnitude of climate change velocity from temperature and precipitation (Supplementary Fig. 19). This combined metric does not include the relative directions of these components.

Residence time. The WDPA is a joint product of the United Nations Environment Programme (UNEP) and International Union for Conservation of Nature (IUCN), prepared by the UNEP World Conservation Monitoring Centre (UNEP-WCMC), supported by the IUCN World Commission on Protected Areas (WCPA) and working with governments, the secretariats of multilateral environmental agreements and collaborating non-governmental organizations. For further information, contact protectedareas@unep-wcmc.org. This database lists centroid coordinates and areas for most protected areas. We used the Geographic Information System (GIS) area field only if documented area

nature

field was blank. For each protected area centroid we sampled the speed and biome to compare with protected area size. We define protected area size as the diameter of a circle of equivalent area to the protected area centred on the centroid of the protected area. This assumes that the protected areas are circular and disconnected. The circular assumption would only be violated if the protected areas had systematically different aspect ratios and orientations by biome. Although some protected areas are not disconnected, we assume that this is a small overall proportion of protected areas.

We plotted histograms of reserve size (km) by biome using biomes from the WWF Terrestrial Ecoregions[30] and centroids of 126,068 protected areas from the WDPA (Supplementary Fig. 20). We assumed reserves were circular and disconnected and computed diameters from their areas. We use these diameters as estimates of reserve size in kilometres.

To estimate uncertainty in the residence time calculations, we propagate the upper and lower uncertainty in the velocity calculations. As before, we plot the histograms of residence time grouped by biome (Supplementary Figs 21 and 22). We log transformed the histograms and report geometric means in the manuscript text to accommodate the skewed protected area size data and residence time distributions.

Comparisons with migration rate studies. The probability distribution function of our temperature-based velocities are consistent with those described previously[21] when uncertainty is accounted for (Supplementary Table 2). We also map the speed of temperature change ($km\,yr^{-1}$) by latitude and compare these values to speeds calculated previously[21] using the Biome3 and MAPSS global vegetation models (Supplementary Fig. 23). Velocities reported previously[21] peak in the high latitudes as opposed to the lower latitudes as found in this study. There are many differences in the approaches used that may account for this contrast. A fundamental difference is that velocities here are based purely on climate gradients, whereas those reported previously are based on the distances between modelled present and future biomes. The approach used previously results in velocities being sensitive to biome size. Another key difference is the much finer spatial resolution of our analysis, which emphasizes the importance of topographic variability. The latter may partially drive the differences in the latitudinal distribution. In this study, higher velocities in lower latitudes are largely driven by the dominance of flat areas (the Amazon Basin and the Sahara, for example) relative to the dominance of steeper topography (and thus lower velocities) in the higher latitudes. The coarser resolution used previously[23] mutes the influence of fine scale topographic variability, contributing to a different latitudinal distribution of climate velocity.

All analyses were performed in R.

31. Daly, C. *et al.* Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.* **28**, 2031–2064 (2008).

nature

LETTERS

A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea

Dongying Wu[1,2], Philip Hugenholtz[1], Konstantinos Mavromatis[1], Rüdiger Pukall[3], Eileen Dalin[1], Natalia N. Ivanova[1], Victor Kunin[1], Lynne Goodwin[4], Martin Wu[5], Brian J. Tindall[3], Sean D. Hooper[1], Amrita Pati[1], Athanasios Lykidis[1], Stefan Spring[3], Iain J. Anderson[1], Patrik D'haeseleer[1,6], Adam Zemla[6], Mitchell Singer[2], Alla Lapidus[1], Matt Nolan[1], Alex Copeland[1], Cliff Han[4], Feng Chen[1], Jan-Fang Cheng[1], Susan Lucas[1], Cheryl Kerfeld[1], Elke Lang[3], Sabine Gronow[3], Patrick Chain[1,4], David Bruce[4], Edward M. Rubin[1], Nikos C. Kyrpides[1], Hans-Peter Klenk[3] & Jonathan A. Eisen[1,2]

Sequencing of bacterial and archaeal genomes has revolutionized our understanding of the many roles played by microorganisms[1]. There are now nearly 1,000 completed bacterial and archaeal genomes available[2], most of which were chosen for sequencing on the basis of their physiology. As a result, the perspective provided by the currently available genomes is limited by a highly biased phylogenetic distribution[3–5]. To explore the value added by choosing microbial genomes for sequencing on the basis of their evolutionary relationships, we have sequenced and analysed the genomes of 56 culturable species of Bacteria and Archaea selected to maximize phylogenetic coverage. Analysis of these genomes demonstrated pronounced benefits (compared to an equivalent set of genomes randomly selected from the existing database) in diverse areas including the reconstruction of phylogenetic history, the discovery of new protein families and biological properties, and the prediction of functions for known genes from other organisms. Our results strongly support the need for systematic 'phylogenomic' efforts to compile a phylogeny-driven 'Genomic Encyclopedia of Bacteria and Archaea' in order to derive maximum knowledge from existing microbial genome data as well as from genome sequences to come.

Since the publication of the first complete bacterial genome, sequencing of the microbial world has accelerated beyond expectations. The inventory of bacterial and archaeal isolates with complete or draft sequences is approaching the two thousand mark[2]. Most of these genome sequences are the product of studies in which one or a few isolates were targeted because of an interest in a specific characteristic of the organism. Although large-scale multi-isolate genome sequencing studies have been performed, they have tended to be focused on particular habitats or on the relatives of specific organisms. This overall lack of broad phylogenetic considerations in the selection of microbial genomes for sequencing, combined with a cultivation bottleneck[6], has led to a strongly biased representation of recognized microbial phylogenetic diversity[3–5]. Although some projects have attempted to correct this (for example, see ref. 5), they have all been small in scope. To evaluate the potential benefits of a more systematic effort, we embarked on a pilot project to sequence approximately 100 genomes selected solely for their phylogenetic novelty: the 'Genomic Encyclopedia of Bacteria and Archaea' (GEBA).

Organisms were selected on the basis of their position in a phylogenetic tree of small subunit (SSU) ribosomal RNA, the best sampled gene from across the tree of life[7]. Working from the root to the tips of the tree, we identified the most divergent lineages that lacked representatives with sequenced genomes (completed or in progress)[8] and for which a species has been formally described[9] and a type strain designated and deposited in a publicly accessible culture collection[10]. From hundreds of candidates, 200 type strains were selected both to obtain broad coverage across Bacteria and Archaea and to perform in-depth sampling of a single phylum. The Gram-positive bacterial phylum *Actinobacteria* was chosen for the latter purpose because of the availability of many phylogenetically and phenotypically diverse cultured strains, and because it had the lowest percentage of sequenced isolates of any phylum (1% versus an average of 2.3%)[11]. Of the 200 targeted isolates, 159 were designated as 'high' priority primarily on the basis of phylum-level novelty and the ability to obtain microgram quantities of high quality DNA. The genomes of these 159 are being sequenced, assembled, annotated (including recommended metadata[12]) and finished, and relevant data are being released through a dedicated Integrated Microbial Genomes database portal[13] and deposited into GenBank. Currently, data from 106 genomes (62 of which are finished) are available.

To assess the ramifications of this tree-based selection of organisms, we focused our analyses on the first 56 genomes for which the shotgun phase of sequencing was completed. The 53 bacteria and 3 archaea (Supplementary Table 1) represent both a broad sampling of bacterial diversity and a deeper sampling of the phylum *Actinobacteria* (26 GEBA genomes). An initial question we addressed was whether selection on the basis of phylogenetic novelty of SSU rRNA genes reliably identifies genomes that are phylogenetically novel on the basis of other criteria. This question arises because it is known that single genes, even SSU rRNA genes, do not perfectly predict genome-wide phylogenetic patterns[14,15]. To investigate this, we created a 'genome tree' (ref. 16) of completed bacterial genomes (Fig. 1) and then measured the relative contribution of the GEBA project using the phylogenetic diversity metric[17]. We found that the 53 GEBA bacteria accounted for 2.8–4.4 times more phylogenetic diversity than randomly sampled subsets of 53 non-GEBA bacterial genomes. A similar degree of improvement in phylogenetic diversity was seen for the more intensively sampled actinobacteria (Table 1). These analyses indicate that although SSU rRNA genes are not a perfect indicator of organismal evolution, their phylogenetic relationships are a sound predictor of phylogenetic novelty within the universal gene core present in bacterial genomes.

[1]DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [2]University of California, Davis, Davis, California 95616, USA. [3]DSMZ, German Collection of Microorganisms and Cell Cultures, 38124 Braunschweig, Germany. [4]DOE Joint Genome Institute-Los Alamos National Laboratory, Los Alamos, California 87545, USA. [5]University of Virginia, Charlottesville, Virginia 22904, USA. [6]Lawrence Livermore National Laboratory, Livermore, California 94550, USA.
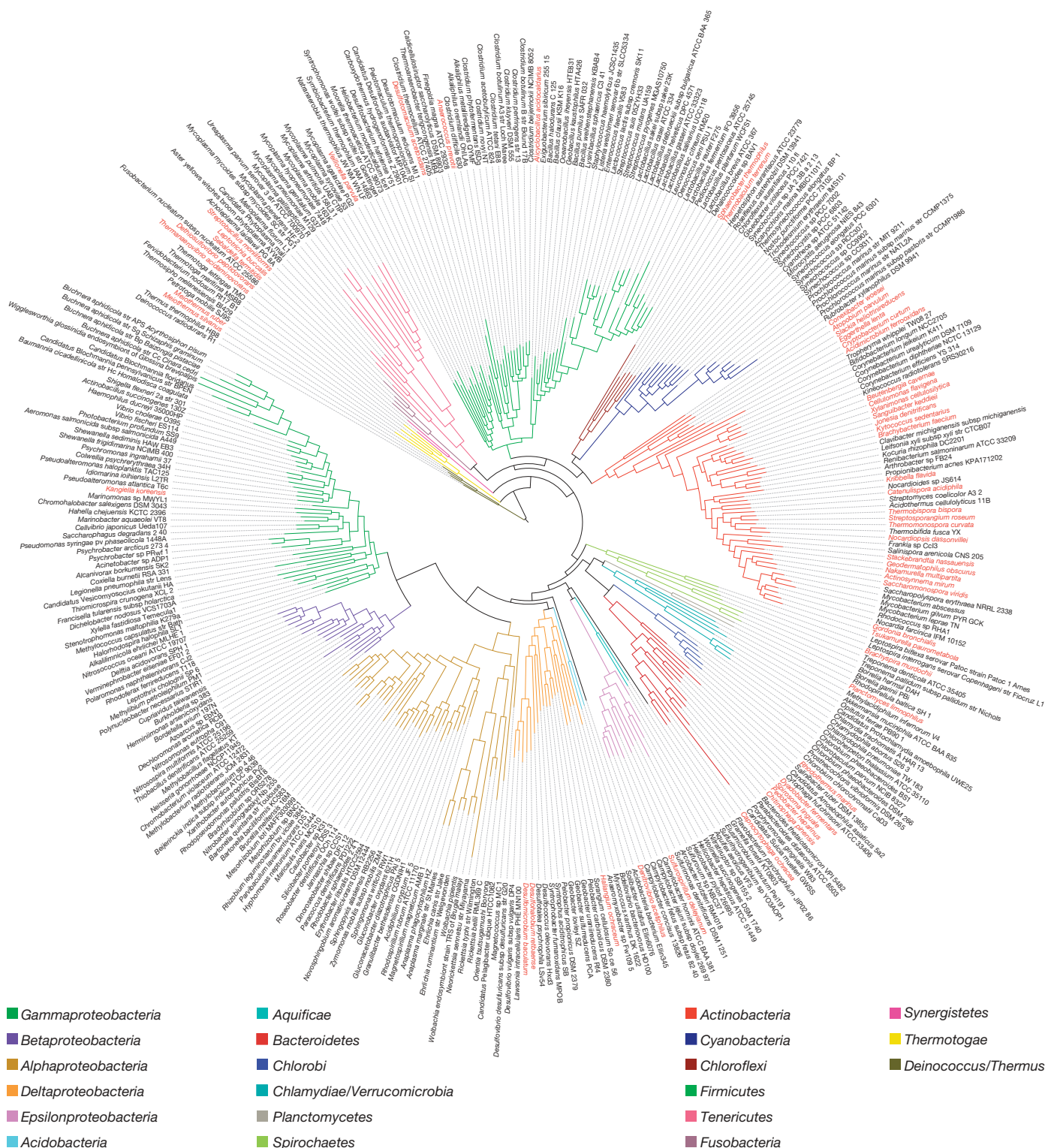
Figure 1 | Maximum-likelihood phylogenetic tree of the bacterial domain based on a concatenated alignment of 31 broadly conserved protein-coding genes[16]. Phyla are distinguished by colour of the branch and GEBA genomes are indicated in red in the outer circle of species names.

Legend:
- Gammaproteobacteria
- Betaproteobacteria
- Alphaproteobacteria
- Deltaproteobacteria
- Epsilonproteobacteria
- Acidobacteria
- Aquificae
- Bacteroidetes
- Chlorobi
- Chlamydiae/Verrucomicrobia
- Planctomycetes
- Spirochaetes
- Actinobacteria
- Cyanobacteria
- Chloroflexi
- Firmicutes
- Tenericutes
- Fusobacteria
- Synergistetes
- Thermotogae
- Deinococcus/Thermus

The discovery and characterization of new gene families and their associated novel functions provide one incentive for sequencing additional genomes, analysis of which has helped to redefine the protein family universe[18]. We explored the quantitative effect of tree-based genome selection on the pace of discovery of novel proteins and functions. Specifically, we compared the rate of discovery of novel protein families when progressively adding more closely related genomes versus when adding more distantly related ones (Fig. 2). Granted, many factors contribute to protein family diversity, such as ecological niche; nevertheless, higher rates of novel protein family discovery were found in the more phylogenetically diverse taxa (Fig. 2). In addition, of the 16,797 families identified in the 56 GEBA genomes, 1,768 showed no significant sequence similarity to any proteins, indicating the presence of novel functional diversity. These results highlight the utility of tree-based genome selection as a means to maximize the identification of novel protein families and argues against lateral gene transfer significantly redistributing genetic novelty between distantly related lineages.

Table 1 | Effect of SSU rRNA tree-based selection of organisms on comparative genomic metrics

Comparative genomic metric	GEBA set	Random sets (number of resamplings)	Fold improvement
Genome tree phylogenetic diversity[17]			
Bacteria (domain)	11.0	3.2 ± 0.7 (100)	2.8–4.4
Actinobacteria (phylum)	4.3	1.4 ± 0.3 (100)	2.5–3.9
New protein family links	46	3 ± 4 (5)	6.6 to >15.3
Genes in new chromosomal cassettes	71,579	16,579 ± 5,523 (20)	3.2–6.5
New gene fusions	433	65 ± 31 (20)	4.5–12.7

GEBA genomes were compared to equivalently sized random sets of reference genomes to quantify the effect of phylogenetic selection.

Novel proteins also can serve to link distantly related homologues whose relatedness would otherwise go undetected. Forty-six such links were identified in the 56 GEBA genomes compared to an average of only three new links in equivalent sets of randomly sampled non-GEBA genomes (Table 1). A useful complement to homology-based predictions of gene function are 'non-homology methods' (ref. 19) such as gene context-based inference that relies on the conserved clustering of functionally related genes across multiple genomes, often in operons or as gene fusions[20]. We identified over 70,000 genes in new chromosomal cassettes of two or more genes in the GEBA genomes. This represents a three- to sixfold increase over equivalent sets of non-GEBA genomes (Table 1). Similarly, the number of new gene fusions identified in the GEBA genomes is 4 to ~13 times greater than in randomly selected genome sets (Table 1). Because the GEBA data set produced a several-fold improvement over random sets for all metrics examined (Table 1), we predict that other aspects of sequence-based biological discovery will similarly benefit from tree-based genome sequencing.

The GEBA genomes also show significant phylogenetic expansions within known protein families. For example, although only two of the 56 GEBA organisms are known cellulose degraders, we identified in the set of genomes a variety of glycoside hydrolase (GH) genes that may participate in the breakdown of cellulose and hemicelluloses. Among these are 28 and 7 phylogenetically divergent members of the endo-glucanase- and processive exoglucanase-containing GH6 and GH48
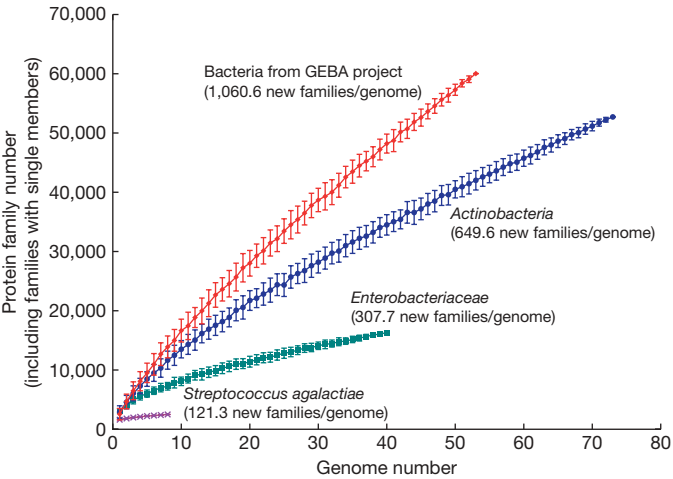
families, respectively. *Halorhabdus utahensis*, a halophilic archaeon known to have β-xylanase and β-xylosidase activities[21], has a chromosomal cluster including two GH10 family β-xylanases and six novel GH5 family proteins of unknown specificity.

The enrichment of genetic diversity is also seen within families of non-coding RNAs, transposable elements, and other cellular components. For example, the genome of the marine myxobacterium *Haliangium ochraceum* contains 807 CRISPR (clustered regularly interspaced short palindromic repeats) units including the largest single CRISPR array known, comprising 382 spacer/repeat units. CRISPR is a newly recognized, but ancient and widespread, system in bacteria and archaea that confers resistance to viruses and other invading foreign DNAs[22].

Results from the GEBA pilot project challenge our current understanding for the taxonomic distribution of known gene families. The most striking example of which is the discovery of an actin homologue in *H. ochraceum*. Actin and its close relatives are structural components of the eukaryotic cytoskeleton that are found in every eukaryote and only in eukaryotes. Bacteria and archaea encode instead the shape-determining protein MreB. Although MreBs have some functional and structural similarities to eukaryotic actins, they are regarded, at best, distantly related homologues[23] and possibly not

Figure 2 | Rate of discovery of protein families as a function of phylogenetic breadth of genomes. For each of four groupings (species, different strains of *Streptococcus agalactiae*; family, *Enterobacteriaceae*; phylum, *Actinobacteria*; domain, GEBA bacteria), all proteins from that group were compared to each other to identify protein families. Then the total number of protein families was calculated as genomes were progressively sampled from the group (starting with one genome until all were sampled). This was done multiple times for each of the four groups using random starting seeds; the average and standard deviation were then plotted.
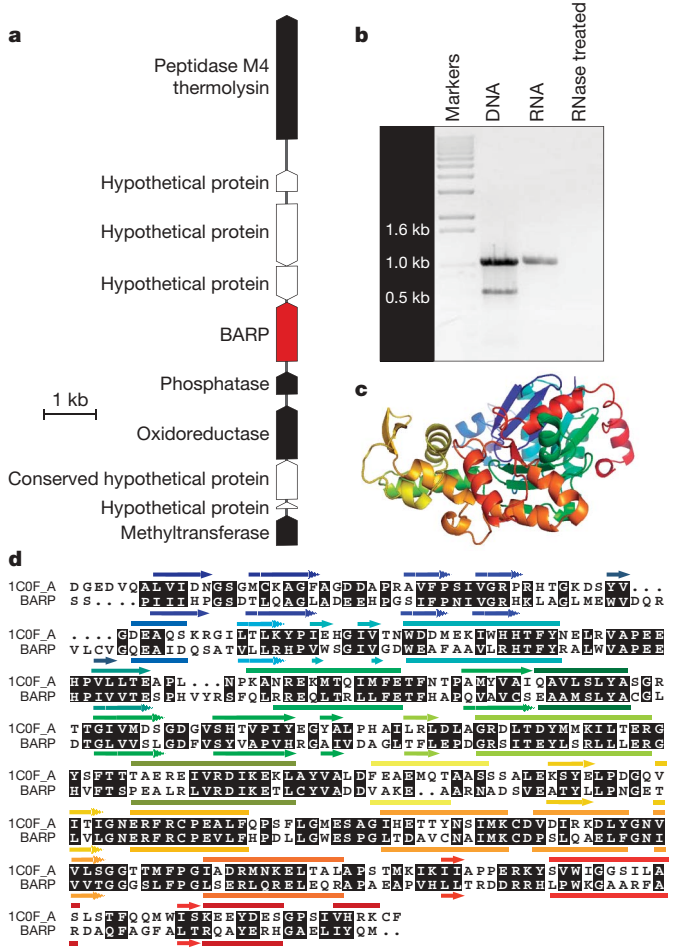
Figure 3 | A bacterial homologue of actin. a, Genomic context of the bacterial actin-related protein (BARP) gene within the genome of the marine Deltaproteobacterium *H. ochraceum*. Red, gene encoding BARP; white, genes encoding hypothetical proteins; black, genes with functional annotations. **b**, RT–PCR demonstration of expression of the gene encoding BARP in *H. ochraceum*. **c**, Ribbon plot of the putative structure of BARP. **d**, Alignment of BARP with actin from *Dictyostelium discoideum*[29] with similarities in black shaded text. Secondary structure elements (arrows, beta-strands; bars, alpha-helices) are colour-coded as in **c**. A phylogenetic tree including this protein is in Supplementary Figure 1.

even homologous. Like other bacteria, *H. ochraceum* encodes a bona fide MreB protein, but in addition, it encodes a protein that is clearly a member of the actin family, which we have named BARP (bacterial actin-related protein; Fig. 3). Although we do not yet have evidence for its precise function, BARP is expressed in *H. ochraceum* (Fig. 3b). Assuming that the *H. ochraceum mreB* orthologue performs the same function as in other bacteria, and given that the myxobacteria, to which this species belongs, are known to synthesize actin-targeting toxins[24], we propose that this BARP may be a dominant-negative inhibitor of eukaryotic actin polymerization. Regardless of its precise function, this first—and so far only—discovery of an expressed homologue of eukaryotic actin in a member of the Bacteria highlights the potential for novel and surprising biological discoveries given a wider genomic sampling of the tree of life.

We conclude that targeting microorganisms for genome sequencing solely on the basis of phylogenetic considerations offers significant far-reaching benefits in diverse areas. Furthermore, the benefits of phylogenetically driven genome sequencing show no sign of saturating with these first 56 genomes. A key question then lies in determining how much bacterial and archaeal diversity remains to be sampled. Using SSU rRNA gene sequences as a proxy for organismal diversity (Fig. 4), we estimate that sequencing the genomes of only 1,520 phylogenetically selected isolates could encompass half of the phylogenetic diversity represented by known cultured bacteria and archaea. Given the continuing reductions in both the cost and difficulty in sequencing genomes[25], this is certainly a tractable target in the next few years.

However, the great majority of recognized bacterial and archaeal diversity is not represented by pure cultures and an additional 9,218 genome sequences from currently uncultured species would be required to capture 50% of this recognized diversity (Fig. 4). Such an undertaking will require new approaches to culturing or processing of multi-species samples using methods such as metagenomics[26] or physical isolation of cells from mixed populations followed by whole genome amplification methods[27]. Obtaining reference genomes for the uncultured microbial majority will be a natural extension of the GEBA project, the ultimate goal of which is to provide a phylogenetically balanced genomic representation of the microbial tree of life. The pilot study presented here is a dedicated first step in this direction.

METHODS SUMMARY

Starting with a phylogenetic tree of SSU rRNA genes[7], we identified major branches that had no available genome sequences but for which cultured isolates were available in the DSMZ or ATCC culture collections. Selected isolates (Supplementary Table 1a, b) from these branches were grown and DNA isolated (Supplementary Table 1c) and quality checked. DNA was then used for shotgun genome sequencing by Sanger/ABI, Roche/454 and/or Illumina/Solexa technologies (Supplementary Table 2). Sequence reads were assembled separately with different assembly methods and the best draft assembly was used for annotation and as a starting point for genome completion (current genome status is in Supplementary Table 2). Annotation (gene identification, functional prediction, etc.) was performed using the IMG system (http://img.jgi.doe.gov/geba); this was done both after shotgun sequencing and again after genome completion. For 'whole genome tree' analysis, a PHYML maximum likelihood phylogenetic tree of a concatenated alignment of 31 marker genes was built using AMPHORA[16]. Phylogenetic diversity was calculated as the sum of branch lengths in this and other trees. Protein families were built for various genome sets by using the Markov clustering algorithm (MCL)[28] to group proteins on the basis of 'all versus all' blastp searches. For analysis of phylogenetic diversity of organisms, a phylogenetic tree was built for a combined alignment of SSU rRNA sequences from published genomes and a non-redundant subset of greengenes SSU rRNA[7]. Further analysis of the genomes was done using IMG database queries and new computational analyses as described in the main text, legends and Supplementary Methods.

1. Fraser, C. M., Eisen, J. A. & Salzberg, S. L. Microbial genome sequencing. *Nature* **406**, 799–803 (2000).
2. Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N. C. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **36** (database issue), D475–D479 (2008).
3. Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**, REVIEWS0003.1–REVIEWS0003.8 (2002).
4. Eisen, J. A. Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr. Opin. Microbiol.* **3**, 475–480 (2000).
5. Wu, D. *et al.* Complete genome sequence of the aerobic CO-oxidizing thermophile *Thermomicrobium roseum*. *PLoS One* **4**, e4207 (2009).
6. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
7. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
8. Bernal, A., Ear, U. & Kyrpides, N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* **29**, 126–127 (2001).
9. Lapage, S. P. *et al.*, International Code of Nomenclature of Bacteria, 1990 Revision. (American Society for Microbiology, 1992).
10. Ward, N., Eisen, J., Fraser, C. & Stackebrandt, E. Sequenced strains must be saved from extinction. *Nature* **414**, 148 (2001).
11. Hugenholtz, P. & Kyrpides, N. C. A changing of the guard. *Environ. Microbiol.* **11**, 551–553 (2009).
12. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol.* **26**, 541–547 (2008).
13. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.* **36** (database issue), D528–D533 (2008).
14. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbiol.* **6**, 431–440 (2008).
15. Beiko, R. G., Doolittle, W. F. & Charlebois, R. L. The impact of reticulate evolution on genome phylogeny. *Syst. Biol.* **57**, 844–856 (2008).
16. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, R151 (2008).
17. Pardi, F. & Goldman, N. Resource-aware taxon selection for maximizing phylogenetic diversity. *Syst. Biol.* **56**, 431–444 (2007).
18. Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V. & Ouzounis, C. A. Myriads of protein families, and still counting. *Genome Biol.* **4**, 401 (2003).
19. Marcotte, E. M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
20. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
21. Wainø, M. & Ingvorsen, K. Production of β-xylanase and β-xylosidase by the extremely halophilic archaeon *Halorhabdus utahensis*. *Extremophiles* **7**, 87–93 (2003).
22. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
23. Doolittle, R. F. & York, A. L. Bacterial actins? An evolutionary perspective. *Bioessays* **24**, 293–296 (2002).
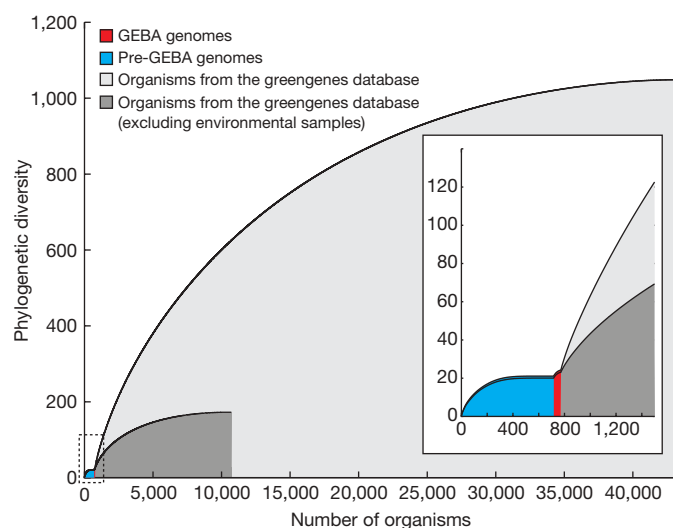

Figure 4 | Phylogenetic diversity of bacteria and archaea on the basis of SSU rRNA genes. Using a phylogenetic tree of unique SSU rRNA gene sequences[7], phylogenetic diversity was measured for four subsets of this tree: organisms with sequenced genomes pre-GEBA (blue), the GEBA organisms (red), all cultured organisms (dark grey), and all available SSU rRNA genes (light grey). For each subtree, taxa were sorted by their contribution to the subtree phylogenetic diversity[30] and the cumulative phylogenetic diversity was plotted from maximal (left) to the least (right). The inset magnifies the first 1,500 organisms. Comparison of the plots shows the phylogenetic 'dark matter' left to be sampled.

24. Sasse, F., Kunze, B., Gronewold, T. M. & Reichenbach, H. The chondramides: cytostatic agents from myxobacteria acting on the actin cytoskeleton. *J. Natl. Cancer Inst.* **90,** 1559–1563 (1998).
25. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnol.* **26,** 1135–1145 (2008).
26. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72,** 557–578 (2008).
27. Ishoey, T., Woyke, T., Stepanauskas, R., Novotny, M. & Lasken, R. S. Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.* **11,** 198–204 (2008).
28. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
29. Matsuura, Y. *et al.* Structural basis for the higher Ca^{2+}-activation of the regulated actin-activated myosin ATPase observed with *Dictyostelium/Tetrahymena* actin chimeras. *J. Mol. Biol.* **296,** 579–595 (2000).
30. Moulton, V., Semple, C. & Steel, M. Optimizing phylogenetic diversity under constraints. *J. Theor. Biol.* **246,** 186–194 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions D.W. (rRNA analysis, gene families, actin tree, manuscript preparation), P.H. (selection of strains, analysis, manuscript preparation, project coordination), L.G. and D.B. (project management), R.P., B.J.T., E.L., S.G., S.S. (strain curation and growth), K.M., N.N.I., I.J.A., S.D.H., A.P., A.Ly. (annotation, genome analysis), V.K. (CRISPRs, actin), M.W. (whole genome tree), P.D., C.K., A.Z. and M.S. (actin studies), M.N., S.L., J.-F.C., F.C. and E.D. (sequencing), C.H., A.La., M.N. and A.C. (finishing), P.C. (analysis), E.M.R. (manuscript preparation), N.C.K. (selection of strains, annotation, analysis), H.-P.K. (strain selection and growth, DNA preparation, manuscript preparation), J.A.E. (project lead and coordination, analysis, manuscript preparation).

LETTERS

Amino-acid imbalance explains extension of lifespan by dietary restriction in *Drosophila*

Richard C. Grandison[1]*, Matthew D. W. Piper[1]* & Linda Partridge[1]

Dietary restriction extends healthy lifespan in diverse organisms and reduces fecundity[1,2]. It is widely assumed to induce adaptive reallocation of nutrients from reproduction to somatic mainten-ance, aiding survival of food shortages in nature[3–6]. If this were the case, long life under dietary restriction and high fecundity under full feeding would be mutually exclusive, through competition for the same limiting nutrients. Here we report a test of this idea in which we identified the nutrients producing the responses of life-span and fecundity to dietary restriction in *Drosophila*. Adding essential amino acids to the dietary restriction condition increased fecundity and decreased lifespan, similar to the effects of full feed-ing, with other nutrients having little or no effect. However, methionine alone was necessary and sufficient to increase fecun-dity as much as did full feeding, but without reducing lifespan. Reallocation of nutrients therefore does not explain the responses to dietary restriction. Lifespan was decreased by the addition of amino acids, with an interaction between methionine and other essential amino acids having a key role. Hence, an imbalance in dietary amino acids away from the ratio optimal for reproduction shortens lifespan during full feeding and limits fecundity during dietary restriction. Reduced activity of the insulin/insulin-like growth factor signalling pathway extends lifespan in diverse organisms[7], and we find that it also protects against the shortening of lifespan with full feeding. In other organisms, including mam-mals, it may be possible to obtain the benefits to lifespan of dietary restriction without incurring a reduction in fecundity, through a suitable balance of nutrients in the diet.

Dietary restriction, whereby food intake is reduced without lead-ing to malnutrition, extends lifespan in many organisms: yeast[8], invertebrates[9] and mammals[1], including primates[10]. In rodents and primates, it also produces a broad-spectrum improvement in health during ageing[1,10]. Reduced calorie intake per se has been suggested to underlie extended lifespan in rodents. However, specific amino acids may be as or more important[11–13]. Dietary restriction lowers fecundity[2], for instance in the nematode worm *Caenorhabditis elegans*[14], the fruit-fly *Drosophila melanogaster*[15] and rodents[16]. The prevailing view is that it induces an evolved response to food shortages[3,17]. If somatic main-tenance and reproduction compete for limiting nutrients then, with abundant food, reproduction is prioritized and somatic maintenance is allocated only the nutrients necessary to ensure survival during the reproductive period, which, owing to extrinsic hazards in the wild, is much shorter than the intrinsic potential lifespan[4]. With food shortage, reproduction becomes dangerous for the parent and offspring survive poorly; nutrients are hence reallocated to somatic maintenance, thus increasing the chances of the organism surviving to reproduce success-fully when the food supply returns[3,5]. In this picture, high survival, associated with dietary restriction, and high reproductive rate, assoc-iated with full feeding, are mutually exclusive.

We have tested this prediction in *Drosophila*. Dietary restriction is implemented by dilution of the diet, without compensation of food intake rates[18–20], resulting in increased lifespan and reduced fecundity, measured as egg laying[20]. In nature, *Drosophila* eat yeasts[21] and, although many manipulations of dietary balance can alter lifespan[22,23], enhanced longevity by dietary restriction is modulated almost exclusively by dietary yeast, independent of calorie intake[18,22–24].

We investigated which nutrients in yeast produce high fecundity in fully fed flies, and whether the same nutrients also decrease lifespan, as predicted by the reallocation hypothesis. The ratio and type of food components were optimized to maximize both lifespan with dietary restriction and fecundity with full feeding[24], and we examined the effect of adding back nutrients to the restricted diet. Because the availability of free nutrients will be higher than that in yeast, we first measured fecundity with addition of all nutrients in the ratio present in yeast (Methods), at several concentrations. We then used the concentration that increased fecundity to the level with full feeding (Supplementary Table 1). Adding back vitamins, lipids or carbohydrates did not affect fecundity or lifespan (Fig. 1), indicating that they do not limit fecundity during dietary restriction and that increased intake of calories per se does not reduce lifespan. In contrast, addition of amino acids increased fecundity and decreased lifespan, as for full feeding (Fig. 1).

To test for non-nutritional toxicity of amino acids, we measured the osmolarity and pH of each diet. In comparison with full feeding, addition of amino acids to the restricted diet caused small changes in osmolarity that do not correlate with lifespan (446 mosM for dietary restriction increased to 495 mosM with the addition of all (essential and non-essential) amino acids and 1,081 mosM for full feeding), and no detectable change in pH, indicating that changes in these factors do not account for the lifespan differences (Supplementary Fig. 2). Furthermore, provision of excess water did not abrogate life-shortening by amino acids, but reduced to zero that caused by the addition of 0.8% salt to the restricted diet (Supplementary Fig. 2), demonstrating the efficacy of water provision.

Reallocation of amino acids from reproduction to somatic main-tenance could explain the responses of lifespan and fecundity to amino acid add-back. Alternatively, different amino acids could indepen-dently produce the two responses. We first investigated the ten essential and ten non-essential amino acids, which are similar in *Drosophila* to those in mammals[25]. Adding back non-essential amino acids (NEAAs) slightly decreased lifespan, with no effect on fecundity (Fig. 2). In contrast, adding back essential amino acids (EAAs) increased fecundity as much as did all 20 amino acids or full feeding (Fig. 2a), and also substantially decreased survival, again as much as full feeding (Fig. 2b). Adding back NEAAs increased dietary nitrogen concentration by 9% more than did adding back EAAs (Supplementary Table 1), suggesting that specific amino acids rather than increased dietary nitrogen were responsible. Further increasing the concentration of EAAs led to
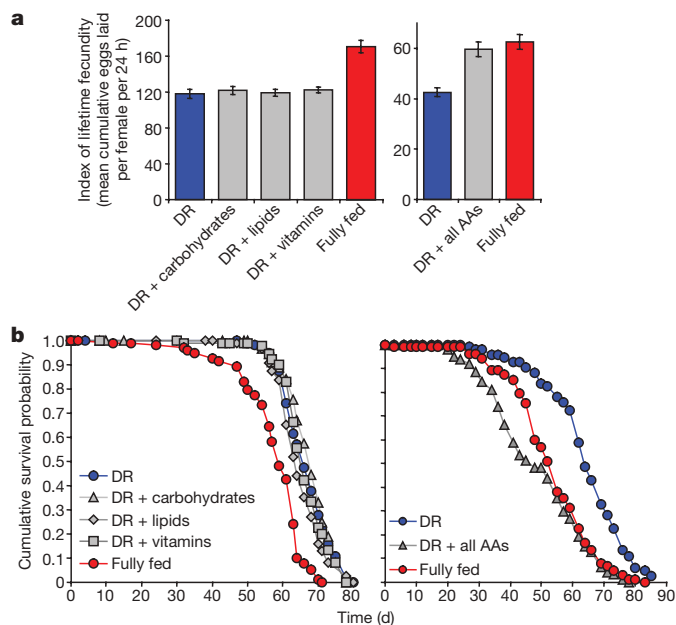
Figure 1 | Amino acids mediate lifespan and fecundity changes in fly dietary restriction. a, Adding nutrients to dietary restriction (DR) revealed that amino acids limited fecundity, because addition of all amino acids (all AAs) restored the fecundity index to that of fully fed flies (DR + all AAs vs fully fed, $P = 0.5288$; DR vs other conditions, $P > 0.2$). **b,** Adding amino acids to DR food shortened lifespan ($P < 0.00001$) to that of fully-fed flies ($P = 0.102$). No other nutrient additions tested changed lifespan from the DR level ($P > 0.1$ in all comparisons). Fecundity: mean ± s.e.m.; $n = 10$; compared using the Wilcoxon rank-sum test. Survivorship: initially $n = 100$ per treatment; compared using the log-rank test.

further increased fecundity and decreased survival (Supplementary Fig. 3). The effects of full feeding can thus be attributed to EAAs in the diet, consistent with reallocation of EAAs from reproduction to somatic maintenance on dietary restriction.

Figure 2 | Essential amino acids cause the dietary restriction effect. a, Adding EAAs, but not NEAAs, increased fecundity to the level reached with all AAs and full feeding (DR + EAAs vs fully fed, $P = 0.393$; DR + EAAs vs DR + all AAs, $P = 1$). **b,** Adding EAAs or all AAs to DR caused lifespan to decrease to the same extent as full feeding ($P > 0.102$). In contrast, NEAA addition to DR shortened lifespan much less ($P = 0.011$). Fecundity (measured as in Fig. 1): mean ± s.e.m.; $n = 10$; compared using the Wilcoxon rank-sum test. Survivorship: initially $n = 100$ per treatment; compared using the log-rank test.

We next determined which EAAs affected fecundity and lifespan. In rodents, lifespan can be extended by restricting either methionine or tryptophan[11–13]. Adding back EAAs except methionine and tryptophan did not increase fecundity (Fig. 3a), indicating that one of these is limiting. Adding back EAAs except methionine also did not increase fecundity from the level corresponding to dietary restriction (Fig. 3a), indicating that methionine is essential, whereas omission of tryptophan produced an increase to the level corresponding to EAA add-back (Fig. 3a). Furthermore, adding back methionine (but not tryptophan or any other EAA) to a restricted diet increased fecundity as much as did the addition of all ten EAAs and full feeding (Fig. 3b and Supplementary Fig. 3). Methionine alone is thus necessary and sufficient for the increase in fecundity. Importantly, egg quality, as indicated by hatching of larvae, was normal with methionine addition (Supplementary Fig. 5). Elevated fecundity with amino-acid addition could have resulted from increased food intake. However, direct feeding observations and dye accumulation assays[19] showed that feeding behaviour and rate of food intake were unaltered (Supplementary Fig. 6). Adding back methionine in a range of concentrations (0.07 mM to 13 mM) increased female fecundity to a plateau (Fig. 3c and Supplementary Fig. 7a), and only addition of other, now limiting, EAAs could increase fecundity further (Supplementary Fig. 8). Thus, methionine probably does not act as a signal to determine fecundity, because its effects depend on the relative concentration of other EAAs, suggesting instead that it acts through nutritional limitation of reproduction.

Unexpectedly, adding back methionine did not decrease lifespan (Fig. 4a), even when it was added back at much higher concentrations
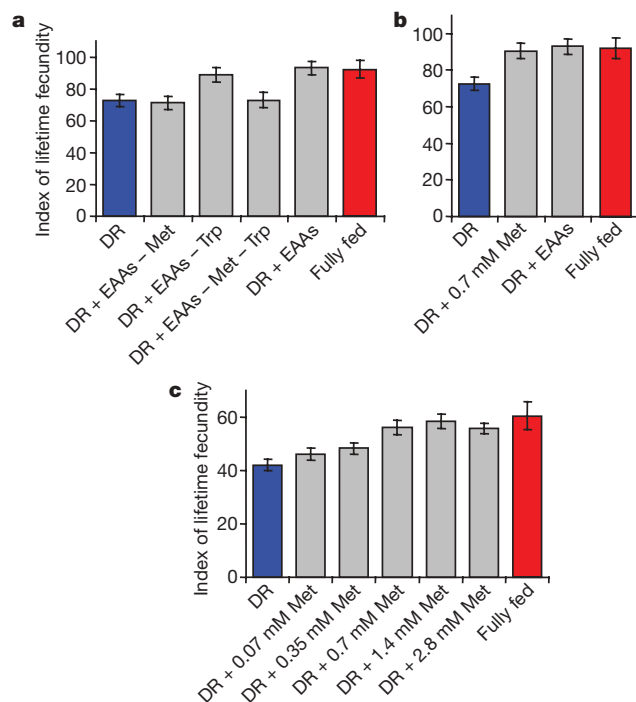

Figure 3 | Methionine is necessary and sufficient to increase dietary restriction fecundity. a, Adding EAAs except methionine (Met) did not increase fecundity from the DR level ($P = 0.796$), whereas adding EAAs without tryptophan (Trp) increased fecundity to the level of DR + EAAs or full feeding (DR vs DR + EAAs − Trp, $P = 0.00893$; DR + EAAs − Trp vs DR + EAAs, $P = 0.4359$; DR + EAAs − Trp vs fully fed, $P = 0.7394$; DR + EAAs − Met − Trp vs DR + EAAs − Met, $P = 0.796$). **b,** Met alone increased fecundity to the level of DR + EAAs and full feeding (DR + Met vs DR + EAAs, $P = 0.5288$; DR + Met vs fully fed, $P = 0.9118$). **c,** Fecundity increased with the addition of Met (significant ≥0.35 mM, $P = 0.02323$), reaching the level of full feeding at 0.7 mM ($P = 0.393$). Fecundity (measured as in Fig. 1): mean ± s.e.m.; $n = 10$; compared using the Wilcoxon rank-sum test.
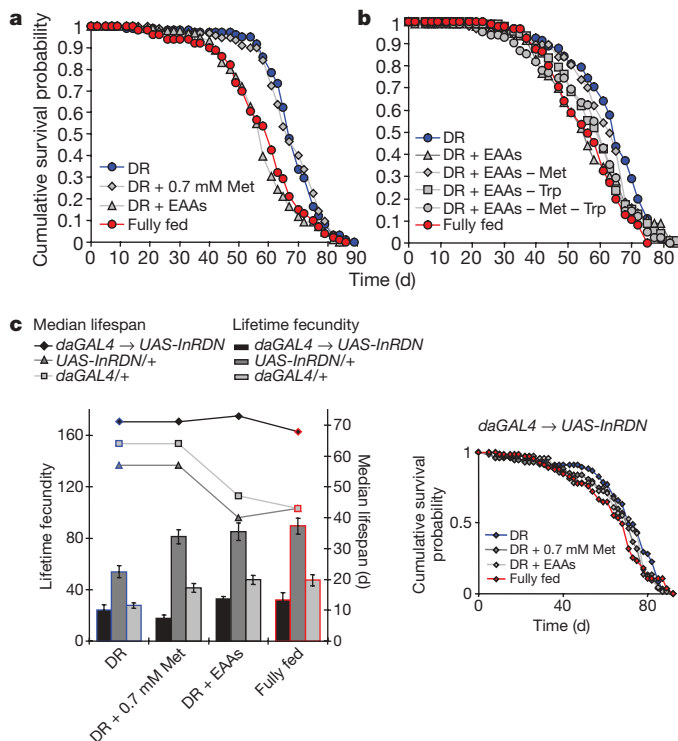
Figure 4 | Amino acids, insulin signalling and dietary restriction.
a, Adding Met to dietary restriction did not shorten lifespan ($P = 0.683$) but adding EAAs did (DR vs DR + EAAs, $P = 0.014$; DR + EAAs vs fully fed, $P = 0.323$). **b**, Adding EAAs without Met restored DR longevity (DR vs DR + EAAs − Met, $P = 0.183$) but Trp removal did not (DR + EAAs − Trp vs fully fed, $P = 0.115$). **c**, Reduced insulin signalling (*daughterless GAL4* (*daGAL4*) → *UAS-InRDN*) extended DR lifespan (DR vs controls, $P < 0.00001$) and reduced flies' response to amino acids or yeast addition (DR vs other foods, $P > 0.064$ (lifespan) and $P > 0.3$ (fecundity)). Lines connect median lifespans. Also shown are survival curves for *daGAL4* → *UAS-InRDN*. Controls responded similarly to wild types. Experimental set-up and statistics as previously described.

than that limiting for fecundity (Supplementary Fig. 7b, c). Hence, reduction in lifespan with full feeding does not result from reallocation of nutrients from somatic maintenance to reproduction, because the nutrient that increased fecundity, methionine, did not reduce lifespan. Furthermore, the fact that high fecundity and high lifespan can co-occur is inconsistent with the idea that any aspect of reproduction directly inflicts damage on the soma to shorten lifespan[26]. We obtained identical results using a fly diet consisting of another yeast commonly used for fly dietary restriction studies[24], indicating that these results are not diet specific (Supplementary Fig. 9). Nor can decreased lifespan with full feeding be attributed to unidentified toxins in the food[20,24]. Instead, the responses of lifespan and fecundity to full feeding are independently mediated by different amino acids.

Adding back each EAA individually did not decrease lifespan, although, again, methionine alone increased fecundity (Supplementary Fig. 4). Adding back all EAAs except methionine restored lifespan to the level corresponding to dietary restriction, whereas omission of tryptophan had no effect (Fig. 4b). Notably, restriction of methionine alone also increases lifespan in rodents[12,13]. Methionine thus acts in combination with one or more other EAAs to shorten lifespan with full feeding. Full feeding thus increases fecundity and decreases lifespan through the effects of different nutrients in *Drosophila*, the fecundity increase through methionine alone and the lifespan decrease through a combination of methionine and other EAAs (Supplementary Fig. 1). There is thus an imbalance in the ratio of amino acids in yeast relative to the ratio the fly requires for the high fecundity from full feeding, and some consequence of this imbalance decreases lifespan.

Genetic interventions that reduce insulin/insulin-like growth factor signalling (IIS) also extend the lifespan of worms, flies and mice[7]. There has been debate on the role of IIS in lifespan extension by dietary restriction[9]. *Drosophila* has a single IIS receptor, InR, which mediates both the growth and the metabolic functions of IIS[27]. We tested the role of IIS in the responses to dietary restriction and EAAs by over-expressing a dominant-negative form of InR, InRDN, that extends fly lifespan[28]. InRDN-expressing flies were longer lived than controls even with dietary restriction and their lifespans, like those of controls, were unchanged by the addition of methionine (Fig. 4c and Supplementary Fig. 10). However, in sharp contrast to controls, lifespan was either only slightly (trial one) or not at all (trial two) reduced by EAA add-back or full feeding (Fig. 4c and Supplementary Fig. 10). InRDN expression also reduced the responses of egg laying to methionine and full feeding. Thus, reduced IIS can both extend lifespan beyond the maximal response to dietary restriction, showing that mechanisms additional to those of dietary restriction are involved, and can also protect against the lifespan-shortening effects of amino-acid imbalance with full feeding and EAA addition, showing that IIS is required for lifespan shortening.

Amino acids that are not used in reproduction in the flies could shorten lifespan through metabolic costs associated with their removal; through consequent damage, for instance to the excretory malpighian tubules; or through other physiological responses. Nutrient imbalance in the diet could also account for the responses of lifespan and fecundity to dietary restriction in other organisms, including mammals, if specific nutrients in their diet are also limiting for full physiological function. Protein quality is implicated in human health, because the ratio of amino acids in the diet can affect traits important for ageing, such as glucose homeostasis and bone health[29]. The mechanisms that influence lifespan are conserved over the large evolutionary distances between invertebrates and mammals[7], and our results hence imply that in mammals also the benefits of dietary restriction for health and lifespan may be obtained without impaired fecundity and without dietary restriction itself, by a suitable balance of nutrients in the diet.

METHODS SUMMARY

Nutritional analysis of the yeast was provided by MP Biomedicals. We prepared stock solutions for the different nutrient add-back treatments as outlined in Supplementary Table 1 and added them to the dietary restriction medium containing $100\,g\,l^{-1}$ yeast, $50\,g\,l^{-1}$ sucrose, $15\,g\,l^{-1}$ agar and preservatives[24], after the food had cooled to 65 °C.

Fly stocks and maintenance. We performed all experiments using the wild-type, outbred strain Dahomey, which was originally collected in 1970 from Dahomey (now the Republic of Benin) and has since been maintained in stock cages with overlapping generations at 25 °C on a 12 h:12 h light–dark cycle. Insulin-signalling mutant flies expressed a dominant negative form of the insulin receptor with a single amino-acid substitution in the kinase domain (UAS-$dInR^{A1409K}$). Adult flies expressing this transgene are approximately 32% lighter than controls, which effect is similar to that of reducing insulin ligand production[30]. To drive ubiquitous expression of the transgene, a *daGAL4* driver was used. Control lines contained either the driver or the UAS transgene alone. All flies were backcrossed into the wild-type w[Dah] background as described[28].

Lifespan and fecundity protocols. We reared flies at standard density and allowed them to mate for 48 h (ref. 24). Under CO_2 anaesthesia, females were collected and randomly allocated to glass vials containing the different add-back treatments, at a density of 10 flies per vial and 10 vials per treatment ($n = 100$). Flies were transferred to fresh medium three times per week and deaths recorded. Egg counts were performed over an 18–24-h period at several intervals throughout the experiment (Methods) until daily egg laying reached a low plateau late in life.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Weindruch, R. & Walford, R. L. *The Retardation of Aging and Disease by Dietary Restriction* (Thomas, 1988).

2. Partridge, L., Gems, D. & Withers, D. J. Sex and death: what is the connection? *Cell* **120**, 461–472 (2005).

3. Holliday, R. Food, reproduction and longevity: is the extended lifespan of calorie-restricted animals an evolutionary adaptation? *Bioessays* **10**, 125–127 (1989).

4. Williams, G. C. Natural selection, the costs of reproduction, and a refinement of Lack's principle. *Am. Nat.* **100**, 687–690 (1966).

5. Kirkwood, T. B. Evolution of ageing. *Nature* **270**, 301–304 (1977).

6. Mair, W. & Dillin, A. Aging and survival: the genetics of life span extension by dietary restriction. *Annu. Rev. Biochem.* **77**, 727–754 (2008).

7. Russell, S. J. & Kahn, C. R. Endocrine regulation of ageing. *Nature Rev. Mol. Cell Biol.* **8**, 681–691 (2007).

8. Kaeberlein, M., Burtner, C. R. & Kennedy, B. K. Recent developments in yeast aging. *PLoS Genet.* **3**, e84 (2007).

9. Piper, M. D. & Bartke, A. Diet and aging. *Cell Metab.* **8**, 99–104 (2008).

10. Colman, R. J. *et al.* Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science* **325**, 201–204 (2009).

11. De Marte, M. L. & Enesco, H. E. Influence of low tryptophan diet on survival and organ growth in mice. *Mech. Ageing Dev.* **36**, 161–171 (1986).

12. Zimmerman, J. A. *et al.* Nutritional control of aging. *Exp. Gerontol.* **38**, 47–52 (2003).

13. Miller, R. A. *et al.* Methionine-deficient diet extends mouse lifespan, slows immune and lens aging, alters glucose, T4, IGF-I and insulin levels, and increases hepatocyte MIF levels and stress resistance. *Aging Cell* **4**, 119–125 (2005).

14. Klass, M. R. Aging in the nematode *Caenorhabditis elegans*: major biological and environmental factors influencing life span. *Mech. Ageing Dev.* **6**, 413–429 (1977).

15. Chapman, T. & Partridge, L. Female fitness in *Drosophila melanogaster*: an interaction between the effect of nutrition and of encounter rate with males. *Proc. R. Soc. Lond. B* **263**, 755–759 (1996).

16. Selesniemi, K., Lee, H. J. & Tilly, J. L. Moderate caloric restriction initiated in rodents during adulthood sustains function of the female reproductive axis into advanced chronological age. *Aging Cell* **7**, 622–629 (2008).

17. Harrison, D. E. & Archer, J. R. Natural selection for extended longevity from food restriction. *Growth Dev. Aging* **53**, 3–6 (1989).

18. Mair, W., Piper, M. D. & Partridge, L. Calories do not explain extension of lifespan by dietary restriction in *Drosophila*. *PLoS Biol.* **7**, e223 (2005).

19. Wong, R. *et al.* Quantification of food intake in *Drosophila*. *PLoS ONE* **4**, e6063 (2009).

20. Piper, M. D. & Partridge, L. Dietary restriction in *Drosophila*: Delayed aging or experimental artefact? *PLoS Genet.* **3**, e57 (2007).

21. Spieth, H. T. Courtship behaviour in *Drosophila*. *Annu. Rev. Entomol.* **19**, 385–405 (1974).

22. Skorupa, D. A. *et al.* Dietary composition specifies consumption, obesity, and lifespan in *Drosophila melanogaster*. *Aging Cell* **7**, 478–490 (2008).

23. Lee, K. P. *et al.* Ageing and reproduction in *Drosophila*: new insights from nutritional geometry. *Proc. Natl Acad. Sci. USA* **105**, 2498–2503 (2008).

24. Bass, T. M. *et al.* Optimization of dietary restriction protocols in *Drosophila*. *J. Gerontol. A* **62**, 1071–1081 (2007).

25. Sang, J. H. & King, R. C. Nutritional requirements of axenically cultured *Drosophila melanogaster* adults. *J. Exp. Biol.* **38**, 793–809 (1961).

26. O'Brien, D. M. *et al.* Use of stable isotopes to examine how dietary restriction extends *Drosophila* lifespan. *Curr. Biol.* **18**, R155–R156 (2008).

27. Fernandez, R. *et al.* The *Drosophila* insulin receptor homolog: a gene essential for embryonic development encodes two receptor isoforms with different signaling potential. *EMBO J.* **14**, 3373–3384 (1995).

28. Ikeya, T. *et al.* The endosymbiont Wolbachia increases insulin/IGF-like signalling in Drosophila. *Proc. R. Soc. B* **276**, 3799–3807 (2009).

29. Millward, D. J. *et al.* Protein quality assessment: impact of expanding understanding of protein and amino acid needs for optimal health. *Am. J. Clin. Nutr.* **87**, 1576S–1081S (2008).

30. Ikeya, T. *et al.* Nutrient-dependent expression of insulin-like peptides from neuroendocrine cells in the CNS contributes to growth regulation in *Drosophila*. *Curr. Biol.* **12**, 1293–1300 (2002).

Author Contributions The project was conceived by M.D.W.P. and L.P., and the experiments were designed by R.C.G., M.D.W.P. and L.P. The experiments were performed and analysed by R.C.G. and M.D.W.P. The manuscript was written by R.C.G., M.D.W.P. and L.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.P. (l.partridge@ucl.ac.uk).

METHODS

Preparation of add-back solutions. To establish the ratios of nutrients present in yeast, we combined data from the literature[31] and chemical analyses[24]. Free nutrients are likely to be at higher effective concentrations than are nutrients in yeast, so we measured the effects on fecundity of an all-nutrient addition, in the ratios in which the individual nutrients are found in yeast (Supplementary Table 1), at several concentrations. We used the concentration that produced the same increase in fecundity as full feeding (Fig. 1a). To check for toxicity, these levels were doubled during NEAA add-back (Fig. 2a, b) and for additions of single EAAs (Supplementary Fig. 4). The individual ingredients were weighed out and dissolved in appropriate solvent to make stock solutions, as outlined in Supplementary Table 1. Phosphatidylcholine was chosen as the lipid source because it is a major phospholipid of eukaryotic cells and contains choline, which is essential for adult *Drosophila*[32]. Food in the restricted diet contains sufficient levels of nutrients from yeast for healthy *Drosophila* lifespan[24]. Amino-acid purity was at least 98%; order numbers are A5131 (Arg), A7627 (Ala), A0884 (Asn), A6683 (Asp), C1276 (Cys), G1251 (Glu), G3126 (Gln), G7126 (Gly), H8000 (His), I2752 (Ile), L8912 (Leu), L5626 (Lys), M9625 (Met), P2126 (Phe), P0380 (Pro), S4500 (Ser), T8625 (Thr), T0254 (Trp), T3754 (Tyr) and V0500 (Val). All add-back reagents were obtained from Sigma.

Media preparation. The restricted diet ($1.0\times$) contained, per litre, 100 g auto-lysed yeast powder (MP Biomedicals), 50 g sucrose (Tate & Lyle Sugars), 15 g agar (Sigma), 30 ml Nipagin ($100\,\mathrm{g\,l^{-1}}$ in 95% ethanol; Clariant) and 3 ml propionic acid (Sigma). This was used as base for all treatments. The full-feeding diet ($2.0\times$) was the same as the restricted diet, except that the autolysed-yeast content was doubled, to $200\,\mathrm{g\,l^{-1}}$. This diet is optimized calorically and nutritionally for dietary restriction experiments with *Drosophila*[24,33]. Separate 1-l quantities were prepared for each treatment. In all cases, the food was prepared as described in ref. 24, and when the temperature fell below 65 °C, the add-back solution was added with any remaining water to adjust the volume, along with preservatives. The food was dispensed into glass vials in 4-ml aliquots. Fresh food was prepared approximately every three weeks throughout the course of the experiments.

Osmolarity determination. A slurry of sugar and yeast at the concentration in the dietary restriction and full-feeding conditions was made in 50 ml water. This was heated to 100 °C and cooled to mimic cooking. A small sample was taken for osmolarity determination using an Advanced Model 3300 Micro-Osmometer (Advanced Instruments).

pH determination. Samples were taken of the food used for experiments. Water (1 ml) was added to the surface of the food and allowed to equilibrate overnight. The water was then removed and the pH measured.

Lifespan experiments. Larvae were reared at a standard density in 200-ml bottles containing 70 ml of 1.0 sucrose–yeast laboratory medium[34]. Flies emerging over a 24-h period were transferred into fresh bottles where they were kept to mate for 48 h. Females were then separated from males under light CO_2 anaesthesia and systematically distributed between food treatments at a density of ten flies per vial, with at least 100 flies per experimental condition. Flies were transferred to fresh vials at least three times per week and deaths scored on those days.

Fecundity assays (index of lifetime fecundity). Eighteen to twenty-four hours after transferring flies to fresh vials, we counted the eggs in each vial by hand under a dissecting microscope. For each vial, the data were expressed as eggs per female per 24 h. At the end of the experiment, the values for each vial were summed to give an index of lifetime fecundity. Typically, the counts were performed on days 6, 8, 15, 22, 29 and 35 after the start of treatment.

Feeding assays. Dye-calibrated feeding observation assays were performed as described in ref. 19. This involves the following two procedures.

(1) Blue-dye feeding assay combined with observations. To determine whether there was a correspondence between observed feeding activity and actual food consumption, the behavioural assay was calibrated. Once mated, female flies were reared and maintained as in the lifespan experiments and housed at five flies per vial. On day seven of adult life, flies were transferred to food containing 2.5% w/w blue dye (FD&C Blue No.1). During the initial 30 min of access to the blue-dyed food, feeding observations were recorded for each of the vials. After 30 min, flies were frozen and the amount of blue dye consumed measured spectrophotometrically. Data were quantified by reference to a standard curve generated from a known amount of blue-dyed food. The relationship between observed feeding events and blue-dye consumption was then analysed. There was a significant linear relationship between the proportion of flies feeding and the amount of dye consumed.

(2) Feeding observations. To measure fly feeding under conditions resembling those in the lifespan experiments, observations of feeding behaviour were made on undisturbed flies. On the days before measurement, flies were transferred to fresh vials, the labelling coded by another lab member and the vials arranged on viewing racks. One hour after lights-on (11 am), feeding observations were made for 90 min as described in ref. 19. Data are presented as the proportion of flies feeding (Supplementary Fig. 6). This is the sum of scored feeding events divided by the total number of feeding opportunities, where the number of feeding opportunities equals the number of flies in the vial multiplied by the number of vials in the group multiplied by the number of observations per vial.

Statistical analyses. All statistical tests were performed using JMP (version 5.1; SAS Institute) and R[35] (version 2.2.1). Survivorships were compared using the log-rank test and fecundities were compared using the Wilcoxon rank-sum test. To assess the relationship between proboscis extension behaviour and accumulation of blue dye, we used a linear mixed-effects model[19]. This modelled dye accumulation as a function of the proportion of time feeding was observed. To compare the effect of dietary composition on feeding frequency, we used generalized linear models (with a binomial error structure and a logit link function, deviances scaled to correct for overdispersion, and *F*-tests for analysis of significance). Simplification of the factor levels was performed by evaluating whether combining factor levels into a single level led to a significant increase in deviance of the model, using *F*-tests[36].

31. Lange, H. C. & Heijnen, J. J. Statistical reconciliation of the elemental and molecular biomass composition of *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **75**, 334–344 (2001).
32. Sang, J. H. in *The Genetics and Biology of Drosophila* (eds Ashburner, M. & Wright, T. R. F.) 159–192 (Academic, 1978).
33. Grandison, R. C. *et al.* Effect of a standardised dietary restriction protocol on multiple laboratory strains of *Drosophila melanogaster*. *PLoS ONE* **4**, e4067 (2009).
34. Clancy, D. J. & Kennington, W. J. A simple method to achieve consistent larval density in bottle cultures. *Drosoph. Inf. Serv.* **84**, 168–169 (2001).
35. R Development Core Team. R: A Language and Environment for Statistical Computing. *The R Project for Statistical Computing* ⟨http://www.R-project.org⟩ (2005).
36. Crawley, M. J. *Statistics: An Introduction Using R* 103–124 (Wiley, 2005).

LETTERS

Secreted semaphorins control spine distribution and morphogenesis in the postnatal CNS

Tracy S. Tran[1,2], Maria E. Rubio[3], Roger L. Clem[1,2], Dontais Johnson[1,2], Lauren Case[4], Marc Tessier-Lavigne[4,5], Richard L. Huganir[1,2], David D. Ginty[1,2] & Alex L. Kolodkin[1,2]

The majority of excitatory synapses in the mammalian CNS (central nervous system) are formed on dendritic spines[1], and spine morphology and distribution are critical for synaptic transmission[2–6], synaptic integration and plasticity[7]. Here, we show that a secreted semaphorin, Sema3F, is a negative regulator of spine development and synaptic structure. Mice with null mutations in genes encoding Sema3F, and its holoreceptor components neuropilin-2 (Npn-2, also known as Nrp2) and plexin A3 (PlexA3, also known as Plxna3), exhibit increased dentate gyrus (DG) granule cell (GC) and cortical layer V pyramidal neuron spine number and size, and also aberrant spine distribution. Moreover, Sema3F promotes loss of spines and excitatory synapses in dissociated neurons *in vitro*, and in $Npn\text{-}2^{-/-}$ brain slices cortical layer V and DG GCs exhibit increased mEPSC (miniature excitatory postsynaptic current) frequency. In contrast, a distinct Sema3A–Npn-1/PlexA4 signalling cascade controls basal dendritic arborization in layer V cortical neurons, but does not influence spine morphogenesis or distribution. These disparate effects of secreted semaphorins are reflected in the restricted dendritic localization of Npn-2 to apical dendrites and of Npn-1 (also known as Nrp1) to all dendrites of cortical pyramidal neurons. Therefore, Sema3F signalling controls spine distribution along select dendritic processes, and distinct secreted semaphorin signalling events orchestrate CNS connectivity through the differential control of spine morphogenesis, synapse formation, and the elaboration of dendritic morphology.

Several axon guidance cues, including class 3 semaphorins (Sema3s), have key roles in synapse formation and function[8–11]. For example, Sema3A promotes the elaboration of dendrite complexity *in vitro*[12,13], may similarly affect dendritic spines[14], and both Sema3A and Sema3F can regulate synaptic transmission in acute brain slices[15,16]. Moreover, $Sema3F^{-/-}$ mutant mice exhibit seizures, and the Sema3F receptor Npn-2 is enriched in the postsynaptic density (PSD)[15]. We address here the *in vivo* roles for these guidance cues and their receptors in synaptogenesis.

Sema3F and its receptor Npn-2 are expressed during synaptogenesis in the hippocampus at postnatal day (P)21 (Supplementary Fig. 1). Npn-2 is enriched in the DG molecular layer, where dendrites of granule cells reside (Supplementary Fig. 1a). Sema3F is strongly expressed in the hilus, along the projection pathways of both supra- and infra-pyramidal axons, and also along entorhinal cortex axons that innervate the DG molecular layers (Supplementary Fig. 1d). Therefore, Sema3F and Npn-2 are expressed in patterns consistent with these proteins directing postnatal hippocampal neural circuit formation.

To assess the involvement of Sema3A and Sema3F in the regulation of dendritic morphology and synaptogenesis we performed Golgi analysis on P14, P21 and adult brains of wild-type mice and

mice harbouring targeted mutations in genes encoding class 3 semaphorins and their receptors. We observed abnormal spine morphology and increased spine number in P21 and adult DG GCs in both $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants (Fig. 1a–c, k and Supplementary Fig. 2h–j, n). Similar fully penetrant and expressive spine morphology defects were observed on apical dendrites of P21 (Supplementary Fig. 2k–m, n) and adult (Fig. 1d–f, j) cortical layer V pyramidal neurons in both $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants. No abnormalities in spine density were observed in $Npn\text{-}2^{-/-}$ mutants at P14 in either DG GCs or layer V neurons (Supplementary Fig. 2d–g, n). Consistent with cortical neuron dendritic spine abnormalities in $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants, we detected endogenous Npn-2 receptors in deep cortical layers and endogenous Sema3F ligand in both the P21 and adult neocortex (Supplementary Fig. 1g, h, j, k), suggesting Sema3F signals through its Npn-2 receptor to regulate spine morphogenesis.

We also observed aberrant spine distribution along apical dendrites of cortical layer V pyramidal neurons in adult $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutant mice; 3.5-fold more spines were present on primary apical dendrites immediately proximal (0–50 μm) to the cell soma, a location where few to no spines were found in wild-type animals (Fig. 1g–j). Analogous spine distribution abnormalities were observed in hippocampal DG GC primary dendrites (Fig. 1k and Supplementary Fig. 2a–c). Spine number normally increases with distance from the soma in wild-type animals[3]. Spine distribution along the middle segments of primary dendrites of both layer V neurons and DG GCs was significantly altered in both $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants (Fig. 1a–f, j, k); however, spine density and morphology along oblique (secondary) branches from primary apical dendrites, and along basal dendrites of layer V cortical neurons, were normal in $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants (Supplementary Fig. 3). Spine number and morphology on both apical and basal dendrites of hippocampal CA1 and cortical layer II/III pyramidal neurons were normal in $Sema3F^{-/-}$ and $Npn\text{-}2^{-/-}$ mutants (Supplementary Fig. 4). Therefore, Sema3F–Npn-2 signalling restricts dendritic spine number, distribution, and regulates spine morphology in select neuronal populations and within distinct dendritic compartments.

We next used a rescue paradigm employing *in utero* electroporation to deliver a Npn-2–IRES–mGFP expression construct to a small number of cortical layer V pyramidal neurons in the $Npn\text{-}2^{-/-}$ mutant cortex at embryonic day 13.5 (E13.5), when deep layer cortical neurons are born. Spine density along apical dendrites was assessed in GFP[+] neurons between P35 and P45. $Npn\text{-}2^{-/-}$ layer V pyramidal neurons expressing mGFP (membrane-localized green fluorescent protein) alone exhibited numerous, aberrant spines along
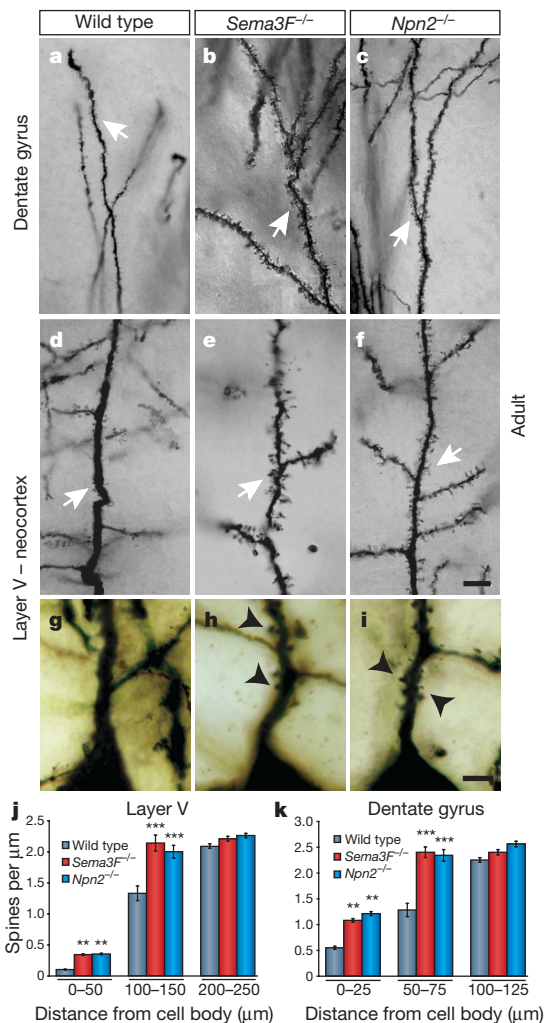
Figure 1 | Sema3F and Npn-2 regulate dendritic spine number, distribution and morphology in adult layer V pyramidal neurons and DG GC in vivo.
a–f, Golgi stained Sema3F$^{-/-}$ (**b**, **e**) and Npn-2$^{-/-}$ (**c**, **f**) brains show that DG GC and layer V pyramidal apical dendritic spines (white arrows) are more numerous compared to wild type (**a**, **d**). **g–i**, Layer V pyramidal neurons have more spines on primary apical dendrites 0–25 μm from the soma in Sema3F$^{-/-}$ (**h**) and Npn-2$^{-/-}$ (**i**) mutants compared to wild-type mice (**g**). **j**, **k**, Quantification of spine density 0–50 μm from the cell body on layer V pyramidal primary apical dendrites (wild type, 0.10 ± 0.01; Sema3F$^{-/-}$, 0.34 ± 0.01; Npn-2$^{-/-}$, 0.35 ± 0.01 spines per μm) and 0–25 μm from the cell body on DG GC primary dendrites (wild type, 0.55 ± 0.03; Sema3F$^{-/-}$, 1.08 ± 0.04 and Npn-2$^{-/-}$, 1.21 ± 0.04 spines per μm). There is a significant increase in spine number on dendritic segments located 100–150 μm from the cell body in layer V (wild type, 1.33 ± 0.14; Sema3F$^{-/-}$, 2.14 ± 0.13; Npn-2$^{-/-}$, 2.00 ± 0.11 spines per μm) and 50–75 μm from the DG neuron cell body (wild type, 1.28 ± 0.13; Sema3F$^{-/-}$, 2.40 ± 0.11; Npn-2$^{-/-}$, 2.34 ± 0.12 spines per μm) in these mutants. There is no significant difference in spine density at 200–250 μm or 100–125 μm from the cell body in layer V and DG neurons, respectively. Error bars, ±s.e.m.; analysis of variance (ANOVA), post-hoc Tukey in **j** and **k**; **$P = 0.01$; ***$P = 0.001$ compared to wild type. Scale bars, 10 μm in **f** for **a–f** and 2.5 μm in **i** for **g–i**.

their apical dendrites (Fig. 2a, b, d). In contrast, Npn-2$^{-/-}$ cortical neurons harbouring the Npn-2–IRES–mGFP construct had 39% fewer spines than Npn-2$^{-/-}$ cortical neurons expressing mGFP alone (Fig. 2c, d). Thus, Npn-2 controls spine number and morphology in a cell-autonomous manner.

To assess the influence of Sema3F on excitatory synapses we treated dissociated cultured wild-type P5 DG neurons with recombinant Sema3F, followed by immunolabelling for the pre- and postsynaptic markers vGlut1 and PSD-95 (also known as Slc17a7 and Dlg4), respectively (Supplementary Fig. 5a–c). Sema3F

Figure 2 | Sema3F–Npn-2 control of spine number is Npn-2 cell-autonomous, and Npn-2 loss-of-function results in increased frequency of mEPSCs. **a–c**, Layer V neurons from P45 wild-type and Npn-2$^{-/-}$ animals from E13.5 embryos in utero electroporated with IRES–mGFP (**a**, **b**), and a Npn-2$^{-/-}$ embryo electroporated with Npn-2–IRES–mGFP (**c**). Scale bars, 10 μm in **c** for **a–c**. **d**, Quantification of spine density (50–100 μm from soma) in Npn-2$^{-/-}$ neurons transfected with Npn-2–IRES–mGFP (0.61 ± 0.03 spines per μm) compared to IRES–mGFP (0.99 ± 0.03 spines per μm), and to wild-type neurons transfected with IRES–mGFP (0.72 ± 0.04 spines per μm). Error bars, ±s.e.m. by ANOVA, post-hoc Tukey. ** and ψ, $P = 0.001$. **e**, Recordings of mEPSCs from cortical slices show a significant increase in mEPSC frequency in Npn-2$^{-/-}$ layer V pyramidal neurons (4.85 ± 0.87 Hz) compared to wild-type littermates (2.04 ± 0.64 Hz). There is no significant difference in mEPSC amplitude between wild-type (9.21 ± 0.32 pA) and Npn-2$^{-/-}$ (8.60 ± 0.35 pA) neurons. Representative mEPSC traces are shown. Error bars, ±s.e.m. by t-test. **$P = 0.024$ for frequency, left panel. $P = 0.231$ for amplitude, right panel. Scale bars, 5 pA (vertical) × 1 s (horizontal).

decreased the average number of puncta exhibiting colocalization of vGlut-1 and PSD-95 by ~50%; however, Sema3A treatment had no effect (Supplementary Fig. 5d). No significant effect on the number of vGlut1 puncta was observed following Sema3F treatment; however, the number of PSD-95-positive puncta decreased by ~40% following Sema3F treatment compared to untreated, or Sema3A-treated, neurons (Supplementary Fig. 5e). The majority of PSD-95-positive puncta in untreated cultured neurons were colocalized with presynaptic vGlut1, and the decrease in the number of excitatory synapses (colocalized vGlut1/PSD-95) was reflected in the decrease in total PSD-95-positive puncta. Sema3F treatment did not affect DG inhibitory synapses (Supplementary Fig. 5f–i). Therefore, Sema3F negatively and selectively influences excitatory

synapses. We performed whole-cell voltage clamp recordings to assess mEPSCs in layer V pyramidal neurons and DG GCs in acute brain slices derived from 3–4-week-old Npn-$2^{-/-}$ and wild-type mice (Fig. 2e and Supplementary Fig. 6). We observed a 2.4- and 1.5-fold increase in mEPSC frequency in Npn-$2^{-/-}$ layer V neurons and DG GCs, respectively, compared to wild-type littermates. Although we observed a slight decrease in the rise time and tau decay for layer V neurons, no significant change in amplitude was observed compared to wild-type littermates (Fig. 2e and Supplementary Fig. 6a). No significant difference in the paired-pulse amplitude ratio was observed between Npn-$2^{-/-}$ and wild-type neurons from layer V or DG (Supplementary Fig. 6), indicating that the increase in mEPSC frequency found in Npn-$2^{-/-}$ mutant mice is due to an increase in the number of synapses rather than an increase in the probability of presynaptic release. These results show that Sema3F–Npn-2 signalling negatively regulates both excitatory synapse number and synaptic transmission in layer V and DG neurons.

To determine how loss of $Sema3F$ and Npn-2 influences synapse formation in $vivo$, we used transmission electron microscopy (TEM) to visualize dendritic spine ultrastructure and excitatory synapse morphology. Spines protruding from the dendritic shafts of wild-type adult DG GCs are small ($<0.1\ \mu m^2$), and of the >270 spines scored (per genotype) we observed that most have round, uniformly shaped, spine heads (Fig. 3a). In contrast, spines in $Sema3F^{-/-}$ and Npn-$2^{-/-}$ mutants are enlarged, vary greatly in shape, and exhibit a >1.7-fold increase in area compared to wild-type spines (Fig. 3a, b and Supplementary Fig. 7a, c, d, f). In spines of mutant mice we observed pre- and postsynaptic components normally associated with wild-type synapses, including electron dense membranous folds in the PSD, vesicle pools near active zones, and docked vesicles associated with presynaptic termini (Fig. 3a, b and Supplementary Fig. 7c, d). However, in adult $Sema3F^{-/-}$ and Npn-$2^{-/-}$ mice we observed a ~5-fold increase in the fraction of DG GC spines harbouring

multiple PSDs (Fig. 3b and Supplementary Fig. 7c, g). Serial section EM reconstructions of several mutant DG GC spines showed that these are perforated PSDs contacted by the same presynaptic terminal (Fig. 3c, d). We found similarly pronounced cortical layer V neuron spine and synaptic morphology defects at the ultrastructural level in both $Sema3F^{-/-}$ and Npn-$2^{-/-}$ mutants (Supplementary Fig. 8). As spine stability, maturation and number increase with age[17,18], we asked whether these abnormalities observed in adult $Sema3F^{-/-}$ and Npn-$2^{-/-}$ mice result from altered spine morphogenesis. Indeed, spines along P21 DG GC dendrites in Npn-$2^{-/-}$ animals already exhibit aberrant morphology similar to that seen in adult Npn-$2^{-/-}$ mutants (Supplementary Fig. 9). These TEM analyses demonstrate that Sema3F and Npn-2 regulate spine morphogenesis and postsynaptic specializations, serving to constrain overall spine number, size and PSD number.

Neuropilins in combination with a class A plexin signalling receptor constitute most secreted semaphorin holoreceptors, and Sema3A preferentially signals through a Npn-1/PlexA4 holoreceptor whereas Sema3F signals through a Npn-2/PlexA3 holoreceptor[19–21]. Analysis of Golgi-labelled adult $PlexA3^{-/-}$ and $PlexA4^{-/-}$ mutant brains revealed that apical dendrite spine morphology is dramatically altered in $PlexA3^{-/-}$, but not $PlexA4^{-/-}$, layer V cortical pyramidal neurons (Fig. 4a, a', b'–d' and Supplementary Fig. 10a), similar to what we observed in $Sema3F^{-/-}$ and Npn-$2^{-/-}$ mutants (Fig. 1d–f). DG GC dendritic spines in both $PlexA3^{-/-}$ and $PlexA4^{-/-}$ mutants are larger, more numerous, and extend much further from the GC
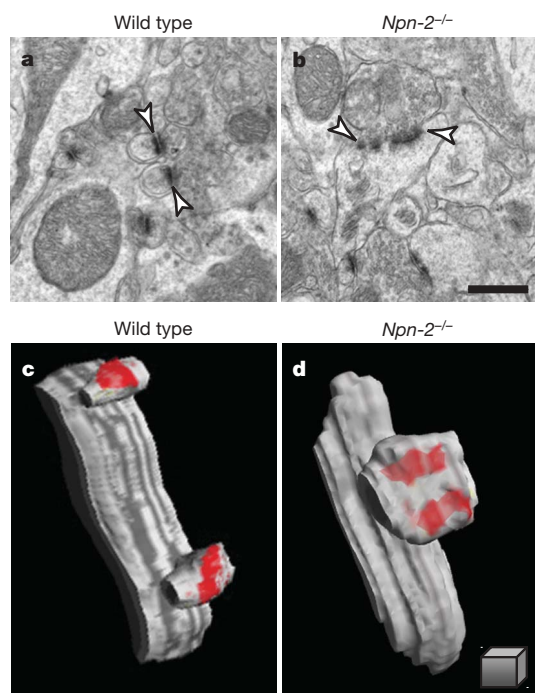

Figure 3 | Sema3F-Npn2 signalling regulates spine morphology and synaptic ultrastructure in vivo. a, b, DG GC dendritic spine TEM ultrastructural analysis shows enlarged and misshapen spines (arrowheads) in $Npn2^{-/-}$ (**b**) compared to wild-type mice (**a**). Scale bars, 500 nm in **b** for **a**, **b**. **c, d**, Three-dimensional reconstructions of serial TEM illustrates two completely separate PSDs within a single spine from a Npn-$2^{-/-}$ mutant mouse (**d**), in contrast to wild-type spines (**c**) with one PSD per spine head. Scale cube, $250 \times 250 \times 250$ nm in **d** for **c, d**.

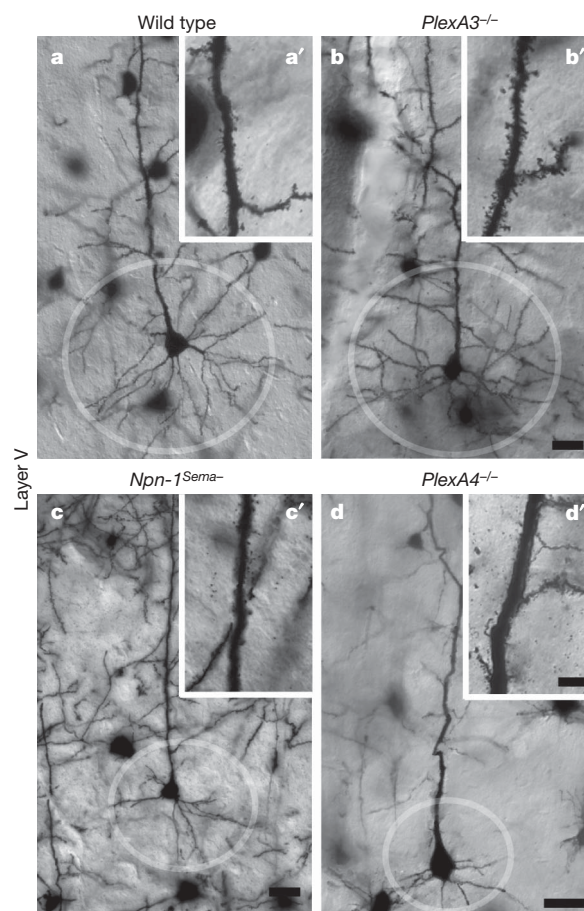**Figure 4 | Distinct Sema3–Npn/PlexA signalling modules regulate apical dendrite spine morphology and basal dendrite process complexity. a–d**, Golgi-labelled adult brains illuminate basal dendritic morphologies in cortical layer V pyramidal neurons from wild-type (**a**, circle), $PlexA3^{-/-}$ (**b**), Npn-1^{Sema-} (**c**), and $PlexA4^{-/-}$ (**d**) mice. **a'–d'**, show spine morphologies from neurons in **a–d**. Scale bars, 10 μm in **d** for **a–d** and 4 μm in **d'** for **a'–d'**.

dendrite shaft than wild-type spines (Supplementary Fig. 11a–d). TEM ultrastructural analysis of $PlexA3^{-/-}$ and $PlexA4^{-/-}$ mice also revealed enlarged and irregularly shaped DG GC spines (Supplementary Fig. 7a, b, e–g). The requirement for PlexA3 and PlexA4 for normal DG GC spine morphology is reminiscent of the requirement for both plexins *in vivo* for correct guidance and extension of embryonic trigeminal neurons, and *in vitro* for repulsive responses to high levels of Sema3A (ref. 21).

Mice homozygous for a knock-in mutation that expresses a Npn-1 protein incapable of binding to Sema3A ($Npn-1^{Sema-}$) phenocopy embryonic neuronal defects observed in *Npn-1* null mice and exhibit markedly reduced growth and branching of layer V cortical neuron basal dendritic arborization[12]. Moreover, acute application of Sema3A to wild-type brain slices promotes an increase in growth and branching of basal dendritic arbors[13]. However, $Npn-1^{Sema-}$ (Fig. 4a, a′, c, c′ and Supplementary Fig. 10a) and $Sema3A^{-/-}$ (not shown) mice do not exhibit spine density defects along apical dendrites of cortical layer V neurons. These results show that Sema3A–Npn-1 signalling positively regulates dendrite growth and branching, but they do not address whether plexin signalling underlies these functions or whether defects in spine morphology in *Sema3F*, *Npn-2* and *PlexA3* mutants are correlated with other dendrite morphogenesis defects. Therefore, we performed Golgi staining on adult brains from wild-type, $Sema3A^{-/-}$, $Sema3F^{-/-}$, $Npn-1^{Sema-}$, $Npn-2^{-/-}$, $PlexA3^{-/-}$ and $PlexA4^{-/-}$ mice. Dendrite orientation and branching in DG GCs in all mutants analysed was identical to wild type (Supplementary Fig. 11e–k). Layer V cortical neuron basal dendrites were also similar to wild type in $Sema3F^{-/-}$, $Npn-2^{-/-}$ and $PlexA3^{-/-}$ mutants (Fig. 4a, b; $Sema3F^{-/-}$ and $Npn-2^{-/-}$ not shown). However, $Npn-1^{Sema-}$, $PlexA4^{-/-}$ and $Sema3A^{-/-}$ mice exhibited severe reductions (~3.8-fold at 63 μm from the cell soma) in the elaboration of basal dendrites of these cortical pyramidal neurons compared to wild type (Fig. 4a, c, d and Supplementary 10b; $Sema3A^{-/-}$ not shown). Therefore, spine defects observed along layer V cortical apical dendrite processes in *Sema3F*, *Npn-2* and *PlexA3* mutant mice are not correlated with the basal dendritic arborization phenotypes observed in the $PlexA4^{-/-}$, $Npn-1^{Sema-}$ and $Sema3A^{-/-}$ mutants[12,13], thus revealing distinct functions of Sema3A–Npn-1/PlexA4 signalling in the promotion of basal dendrite complexity, and of Sema3F–Npn-2/PlexA3 in constraining spine number, distribution and synaptic transmission.

To ask how distinct Sema3 signalling pathways independently regulate cortical neuron basal dendrite morphology and apical dendrite spine morphogenesis, we examined subcellular Npn receptor distribution on dissociated E14.5 cortical neurons grown for 18 days *in vitro* using alkaline phosphatase (AP)-Sema3F (to reveal cell surface Npn-2) and AP–Sema3A (to reveal Npn-1). We observed that cell surface Npn-2 receptors were predominantly localized to the primary apical dendrite in cortical neurons with pyramidal morphology (Fig. 5a–f, j); Npn-2 was absent from both basal dendrites and oblique, or secondary, branches off of primary apical dendrites in these neurons (Fig. 5a, d, k and Supplementary Fig. 12b). This observation was confirmed using Npn-2 antibodies (Supplementary Fig. 12a–a′′′). Npn-1 receptors, in contrast, were more uniformly distributed on all dendritic processes (Fig. 5g–j). This exquisite pattern of Npn-2 distribution likely explains the restricted effects of Sema3F on spines associated with primary apical dendrites (Fig. 1j and Supplementary Fig. 3).

To ask whether Sema3F–Npn-2 signalling can directly regulate cortical neuron spine density and morphology, we next employed an assay in which dendritic spines are visualized following transfection of primary neuronal cultures derived from mouse E14.5 neocortex with an IRES-myristoylated GFP (mGFP) construct (Supplementary Fig. 13a). Sema3F-treated wild-type cortical neurons with pyramidal morphology have along their apical, but not basal, dendrites 33% fewer spines than do untreated (control) or Sema3A-treated neurons (Supplementary Fig. 13b–c, j). In contrast, $Npn-2^{-/-}$ cortical neurons transfected with IRES–mGFP displayed a
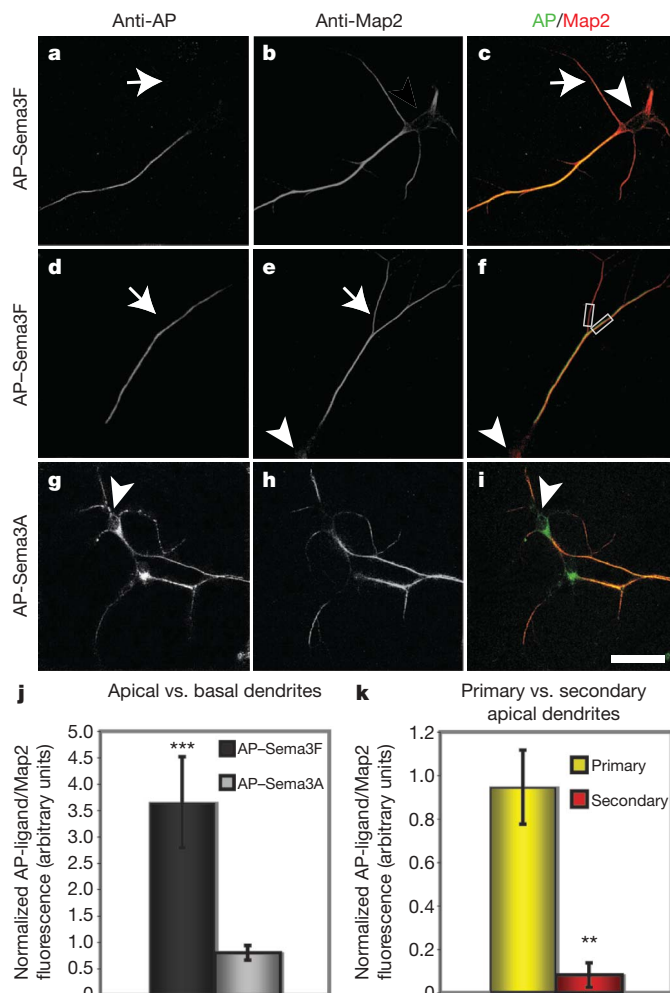

Figure 5 | Npn-1 and Npn-2 are localized to distinct cortical pyramidal neuron dendritic domains. a–i, AP-Sema3F (**a–f**) or AP–Sema3A (**g–i**) fusion proteins were used to localize endogenous Npn-2 and Npn-1, respectively, on cortical pyramidal neurons in primary culture. Endogenous Npn-2 (**a, c, f**) is predominantly restricted to the major apical dendrites of pyramidal neurons as shown by anti-AP labelling (**a–c**). Npn-2 is not observed on basal (white arrows in **a, c**) or secondary apical dendritic branches (white arrows in **d, e**) illuminated by anti-Map2 staining. Npn-1 (**g, i**) is observed on basal and apical dendritic processes (**h, i**). **j, k,** Fluorescence intensities were quantified by measuring apical/basal fluorescence ratios and normalizing to apical/basal Map2 fluorescence ratios for AP–Sema3F-labelled neurons (**j**, 3.70 ± 0.87), and for AP–Sema3A-labelled neurons (0.80 ± 0.14). The Npn-2/Map2 fluorescence ratio in primary apical dendrites (yellow bar, **k**) is 0.95 (± 0.17), and in secondary apical dendrites (red bar, **k**) is 0.08 (± 0.06). Grey boxes, **f**, indicate area of measurement in **k**. Scale bar, 25 μm in **i** for **a–i**.

similar number of spines along their apical dendritic process in both untreated and Sema3F-treated cultures (Supplementary Fig. 13d, e, j). However, $Npn-2^{-/-}$ cortical neurons transfected with a Npn-2–IRES–mGFP construct and subsequently treated with Sema3F have 24% fewer spines along their apical dendrites compared to untreated $Npn-2^{-/-}$ neurons (Supplementary Fig. 13f, g, j). Overexpression of Npn-2 using Npn-2–IRES–mGFP in wild-type dissociated cortical neurons produced no difference in apical or basal dendritic spine number, compared to wild-type neurons transfected with IRES–mGFP (Supplementary Fig. 14a, c, e). However, Sema3F treatment of wild-type neurons overexpressing Npn-2 led to a 30% and 23% reduction in apical and basal dendritic spines, respectively, compared to untreated neurons (Supplementary Fig. 13b, d, e).

The Npn-2 intracellular domain contains a putative C-terminal PDZ ligand motif (SEA) that may be critical for Npn-2/PlexA3 localization

and Sema3F/Npn-2-mediated regulation of spine morphology and synapse structure. We transfected dissociated $Npn-2^{-/-}$ cortical neurons with a Npn-2 SEA-deletion expression construct (Npn-2–ΔSEA–IRES–mGFP) and assessed Sema3F effects on spine morphology. Npn-2–IRES–mGFP and Npn-2–ΔSEA–IRES–mGFP constructs both promote expression of Npn-2 protein on the cell surface of $Npn-2^{-/-}$ cortical neurons (Supplementary Fig. 12d, e). However, $Npn-2^{-/-}$ neurons transfected with Npn-2–ΔSEA–IRES–mGFP do not exhibit a reduction in spine density following Sema3F treatment (Supplementary Fig. 13h–j). Therefore, Sema3F-mediated reduction in spine number along cortical dendritic processes is dependent upon the Npn-2 cytoplasmic SEA PDZ domain-binding motif.

We demonstrate here that spatially segregated secreted semaphorin signalling orchestrates the elaboration of distinct morphological features in select hippocampal and cortical pyramidal neuron dendrites. The organization and distribution of excitatory synapses along primary, secondary and higher order dendritic branches defines how presynaptic inputs are integrated into neural networks. Thus, the precise control of both excitatory and inhibitory synapse distribution during neural development is essential for the formation of functional circuits. Our finding that Sema3F controls the spatial distribution of spines along apical dendrites of cortical pyramidal and hippocampal granule neurons indicates that this secreted cue is essential for integration of excitatory inputs onto these neurons. Supporting this idea, both cortical and hippocampal neurons from $Npn-2^{-/-}$ mutant mice exhibit an increased mESPC frequency, and we observed that mice lacking components of the Sema3F–Npn-2/PlexA3 signalling module exhibit alterations in synaptic transmission (present study) and seizures[15]. These findings underscore the necessity of understanding the mechanisms underlying Sema3F–Npn-2/PlexA3 control of differential spine growth and distribution, and Sema3A–Npn-1/PlexA4 control of basal dendrite growth. Npn-2 is localized to the PSD[15] and we show here that the Npn-2 PDZ domain-binding motif is essential for Sema3F responsiveness. The precise localization of Sema3 holoreceptor complexes via one or more PDZ scaffold proteins associated with postsynaptic components may serve to provide directed Sema3 signalling to subcellular dendritic compartments, regulating dendritic spine morphology and spatial distribution of synapses.

METHODS SUMMARY

The day when a vaginal plug was observed is designated as embryonic day (E) 0.5 and the day of birth as postnatal (P) day 0. Golgi labelling was performed as described (FD NeuroTechnologies). For in utero electroporation, E13.5 embryos from timed-pregnant wild-type and $Npn-2^{-/-}$ females from homozygous crosses were used. Wild-type P5 dentate gyrus hippocampi, and wild-type or $Npn-2^{-/-}$ E14.5 cortices, were dissected and dissociated for primary culture experiments. See Methods for additional experimental procedures, including TEM, serial three-dimensional reconstruction, immunocytochemistry, physiological recordings and quantification parameters for spine density, synaptic puncta, Sholl analysis and AP-fluorescence.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Nimchinsky, E. A., Sabatini, B. & Svoboda, K. Structure and function of dendritic spines. *Annu. Rev. Physiol.* **64**, 313–353 (2002).

2. Elston, G. N. & DeFelipe, J. Spine distribution in cortical pyramidal cells: a common organizational principle across species. *Prog. Brain Res.* **136**, 109–133 (2002).

3. Ballesteros-Yáñez, I., Benavides-Piccione, R., Elston, G. N., Yuste, R. & DeFelipe, J. Density and morphology of dendritic spines in mouse neocortex. *Neuroscience* **138**, 403–409 (2006).

4. Hayashi, Y. & Majewska, A. Dendritic spine geometry: functional implication and regulation. *Neuron* **46**, 529–532 (2005).

5. Alvarez, V. A. & Sabatini, B. L. Anatomical and physiological plasticity of dendritic spines. *Annu. Rev. Neurosci.* **30**, 79–97 (2007).

6. Gao, W.-J. & Zheng, Z.-H. Target-specific differences in somatodendritic morphology of layer V pyramidal neurons in rat motor cortex. *J. Comp. Neurol.* **476**, 174–185 (2004).

7. Spruston, N. Pyramidal neurons: dendritic structure and synaptic integration. *Nature Rev. Neurosci.* **9**, 206–221 (2008).

8. McAllister, A. K. Dynamic aspects of CNS synapse formation. *Annu. Rev. Neurosci.* **30**, 425–450 (2007).

9. Packard, M. et al. The *Drosophila* Wnt, wingless, provides an essential signal for pre- and postsynaptic differentiation. *Cell* **111**, 319–330 (2002).

10. Poon, V. Y., Klassen, M. P. & Shen, K. UNC-6/netrin and its receptor UNC-5 locally exclude presynaptic components from dendrites. *Nature* **455**, 669–673 (2008).

11. Pasterkamp, R. J. & Giger, R. J. Semaphorin function in neural plasticity and disease. *Curr. Opin. Neurobiol.* **19**, 263–274 (2009).

12. Gu, C. et al. Neuropilin-1 conveys semaphorin and VEGF signaling during neural and cardiovascular development. *Dev. Cell* **5**, 45–57 (2003).

13. Fenstermaker, V., Chen, Y., Ghosh, A. & Yuste, R. Regulation of dendritic length and branching by semaphorin 3A. *J. Neurobiol.* **58**, 403–412 (2004).

14. Yamashita, N. et al. Regulation of spine development by Semaphorin3A through cyclin-dependent kinase 5 phosphorylation of collapsin response mediator protein 1. *J. Neurosci.* **27**, 12546–12554 (2007).

15. Sahay, A. et al. Secreted semaphorins modulate synaptic transmission in the adult hippocampus. *J. Neurosci.* **25**, 3613–3620 (2005).

16. Bouzioukh, F. et al. Semaphorin3A regulates synaptic function of differentiated hippocampal neurons. *Eur. J. Neurosci.* **23**, 2247–2254 (2006).

17. Yuste, R. & Bonhoeffer, T. Genesis of dendritic spines: insights from ultrastructural and imaging studies. *Nature Rev. Neurosci.* **5**, 24–34 (2004).

18. Zuo, Y., Lin, A., Chang, P. & Gan, W. Development of long-term dendritic spine stability in diverse regions of cerebral cortex. *Neuron* **46**, 181–189 (2005).

19. Cheng, H.-J. et al. Plexin-A3 mediates semaphorin signaling and regulates the development of hippocampal axonal projections. *Neuron* **32**, 249–263 (2001).

20. Suto, F. et al. Plexin-A4 mediates axon-repulsive activities of both secreted and transmembrane semaphorins and plays roles in nerve fiber guidance. *J. Neurosci.* **25**, 3628–3637 (2005).

21. Yaron, A., Huang, P. H., Cheng, H. J. & Tessier-Lavigne, M. Differential requirement for Plexin-A3 and -A4 in mediating responses of sensory and sympathetic neurons to distinct class 3 Semaphorins. *Neuron* **45**, 513–523 (2005).

Author Contributions T.S.T. performed most of the experiments and data analysis, and M.E.R. performed most of the non-serial TEM analysis. R.L.C. and R.L.H. performed and analysed whole-cell patch recordings. L.C. and M.T.-L. participated in the analysis of *Plexin* mutant mice. D.J. provided technical support. T.S.T., D.D.G. and A.L.K. designed the experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.D.G. (dginty@jhmi.edu) and A.L.K. (kolodkin@jhmi.edu).

METHODS

Golgi Staining. Freshly dissected P14, P21 and adult (3–6-months old) mouse brains were incubated in Golgi solution A+B (FD Rapid GolgiStain Kit, FD NeuroTechnologies) for 7 days. After incubation all brains were washed thoroughly with Solution C for 8 h at room temperature, and then mouse brains were blocked and embedded in OCT embedding medium (Tissue-Tek). Sagittal sections (100 µm) through the medial somatosensory cortex and medial dentate gyrus were cut with a Leica CM3050 cryostat and mounted on 3% gelatin-coated slides. Staining procedures were followed as described (FD NeuroTechnologies), and slides were dehydrated in ethanol and mounted with Permount (Fisher Scientific) for microscopy. Only layer II–III and V pyramidal neurons from the medial somatosensory cortex, hippocampal pyramidal neurons from dorso-medial CA1, and dorso-medial dentate gyrus GNs were included in our analyses.

In utero **electroporation.** All pregnant females were deeply anaesthetized with 10% nembutal via intraperitoneal injection. Surgical procedures to expose the uterus for electroporation of embryos were performed as described[22]. A glass microcapillary pipette was used to deliver 1–2 µg µl^{-1} total DNA into the lateral ventricle of E13.5 embryos. Five electric pulses at 30 V, each pulse with a duration of 50 ms, were delivered at 1-s intervals to each injected embryo. A CD1 foster mother was used to rear all pups until P35–45, when animals were euthanized and analysed. Only anti-GFP-labelled layer V pyramidal neurons from the medial somatosensory cortex were scored. We confirmed the identity of all GFP-labelled cortical neurons by co-labelling with the transcription factor Ctip2, which is expressed mainly in layer V pyramidal neurons[23] (data not shown).

Primary neuronal cultures and Sema3F treatments. All hippocampi and cortices were dissected in ice cold L-15 Leibovitz Medium (Sigma) and incubated in trypsin-EDTA (0.05%; Gibco) at 37 °C for 20–25 min. The tissues were washed twice at room temperature in Ca^{2+}- and Mg^{2+}-free Hank's balanced salt solution (HBSS; Gibco) and dissociated with a fire-polished glass Pasteur pipette in HBSS containing 0.025% DNase I. Dissociated cells were washed twice in HBSS and resuspended in Mouse Neuron Nucleofector Solution (Lonza) for DNA transfections. Following transfection, cells were plated on 12 poly-D-lysine-coated 18-mm glass coverslips (Sigma) and grown in a 12-well dish for 24 h in Opti-MEM (Gibco) containing 10% heated-inactivated horse serum (Invitrogen), 50 U ml^{-1} penicillin, 50 µg ml^{-1} streptomycin (Gibco), 2 mM GlutaMAX (Gibco) and 1% N-2 supplement (Gibco) at 37 °C. Then, primary cultures were switched to Neurobasal medium (Gibco) containing 2% B27 supplement (Gibco) and the same concentrations of GlutaMAX, penicillin and streptomycin as indicated above, and grown for 18–21 days with media changes every other day on glass coverslips. Wild-type and $Npn-2^{-/-}$ primary cultures (transfected with mGFP) were treated with medium containing 5 nM AP–Sema3F, AP–Sema3A (wild-type neurons only) or AP alone (as a control) for either 2 or 6 h (combination of 36 neurons scored for each treatment from $n = 4$ independent cultures; no significant differences in spine number were observed between the 2 and 6 h treatments), fixed in 4% paraformaldehyde at 4 °C for 15 min, and processed for immunocytochemistry. For $Npn-2^{-/-}$ neurons transfected with either Npn2–mGFP or Npn2–ΔSEA–mGFP, and wild-type neurons transfected with either mGFP or Npn2–mGFP, 6 h treatments were performed and a total of 30 neurons were scored from three independent cultures. Basal dendrite spine number quantification was derived from an average obtained by measuring spine density along two separate branches originating from the soma.

Immunocytochemistry. Primary neurons were transfected with the pCAG–IRES (internal ribosome entry site) plasmid, which includes a green fluorescent protein (GFP) containing the amino acid sequence MGCIKSKGKDS (Lyn$_{11}$; ref. 24) that serves to target the protein to the plasma membrane (pCAG–IRES–mGFP), and either a wild-type Npn2 (pCAG–Npn2–IRES–mGFP) or a PDZ ligand-binding motif Npn-2 deletion sequence (pCAG–Npn2–ΔSEA–IRES–mGFP). Dendrite and spine morphologies from primary cultured neurons or from cryosections were visualized using either rabbit (1:500; Molecular Probes) or chicken (1:1,000; Aves Labs) polyclonal antibodies directed against GFP. Glass coverslips upon which the primary neurons were grown were incubated with 0.1% Triton X-100 for 10 min and blocked with TBS-NaCl buffer containing 5% normal goat serum for 1 h at room temperature. The coverslips were then incubated with primary antibodies overnight at 4 °C and subsequently incubated in secondary antibodies for 2 h at room temperature. The following secondary fluorescent antibodies raised in goat were used: anti-mouse, anti-rabbit, anti-chicken or anti-guinea pig IgG conjugated with AlexaFluor 488 (1:200; Invitrogen); anti-mouse, anti-rabbit or anti-guinea pig IgG conjugated with AlexaFluor 546 (1:100; Invitrogen); anti-rat IgG–AlexaFluor 555 (1:100; Invitrogen); and anti-mouse or rabbit IgG conjugated with AlexaFluor 635 (1:100; Invitrogen). Coverslips were mounted using Vectashield hard set fluorescence mounting medium (Vector Laboratories), and confocal stack images

were taken using a Zeiss Axioskop2 Mot Plus, LSM 5 Pascal confocal microscope. The following monoclonal primary antibodies were used: rabbit anti-Npn-2 (1:500; Cell Signaling), mouse anti-PSD-95 (1:100; Chemicon), mouse anti-gephyrin (1:500; Synaptic Systems) and rat anti-Ctip2 (1:500; Abcam). The following polyclonal primary antibodies were also used: guinea pig anti-vGlut1 (1:1000; Chemicon), rabbit anti-vGAT (1:1000, Synaptic Systems), and chicken anti-Map2 (1:1000; Chemicon).

Transmission electron microscopy. Adult (3–6-months-old) and P21 mice were fixed by transcardial perfusion with 3.0% formaldehyde/1.5% glutaraldehyde in 0.1 M Na$^+$-cacodylate, 5 mM Ca^{2+} and 2.5% sucrose, at pH 7.4. Brains were dissected and washed with 0.1 M Na$^+$-cacodylate, 2.5% sucrose, and 200-µm-thick vibratome sagittal sections were collected in 0.1 M phosphate buffer. Medial dentate gyri and medial somatosensory cortices were trimmed, reduced with 1.0% OsO$_4$, 1.0% Kellenberger's uranyl acetate, dehydrated in ethanol, infiltrated and flat-embedded in EPON. Semi-thin and ultra-thin sections were collected using a Leica Ultracut microtome and analysed with a TECNAI G2 Spirit Biotwin TEM (FEI). Images were captured with an AMT XR40 4-megapixel side-mounted CCD camera (Danvers) at ×68,000 magnification. Only identified synapses on dendritic spines of DG granule cells and layer V pyramidal neurons were included in these analysis. To determine spine area, we traced the plasma membrane with ImageJ (http://rsb.info.nih.gov/ij). The number of PSDs per spine was counted manually. Results were compared to wild-type controls by ANOVA, Student's *t*-test and Chi-square (assuming equal or unequal variance according to different data characteristics; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$). Only dendritic spines of layer V and dentate granule neurons from the medial somatosensory cortex and medial dentate gyrus were included in our analyses. Total number of spines scored for area analysis: 277 for wild type, $Sema3F^{-/-}$ and $Npn-2^{-/-}$; 346 for $PlexA3^{-/-}$ and $PlexA4^{-/-}$. Total number of spines scored for PSD/spine analysis: 756 for wild type, 635 for $Sema3F^{-/-}$, 745 for $Npn-2^{-/-}$, 703 for $PlexA3^{-/-}$ and 721 for $PlexA4^{-/-}$.

Three-dimensional serial section electron microscopy. Ultrathin (80 nm) compression free serial sections were cut with a Reichert Ultracut E microtome using a Diatome diamond knife (35 degree). Consecutive sections were picked up on 1 mm × 2 mm Formvar-coated copper slot grids (Polysciences) and stained with uranyl acetate followed by lead citrate. Grids were examined on a Hitachi H-7600 TEM operating at 80 Kv. Images were digitally captured with an AMT 1K × 1K CCD camera at ×30,000 magnification. Dendritic segments from the medial DG granule neurons were traced in Adobe Photoshop CS3 and reconstructed with DeltaViewer 2.1.1 (http://vivaldi.ics.nara-wu.ac.jp/~wada/DeltaViewer/). A total of three stacks of serial images were reconstructed per genotype from $n = 2$ animals per genotype.

Quantification of spine density and synaptic puncta. Spine density (spines per µm) along Golgi-stained neurons in P14, P21 and adult brains were viewed in sagittal sections and quantified using the following parameters: immediately proximal to the soma was defined as the first 50 µm or 25 µm from the cell body for cortical layer V and DGs, respectively. The number of spines was scored by counting spines along a 20 µm segment, or a 10 µm segment, of proximal cortical or DG dendrite, respectively. The middle regions were defined as 100–150 µm for layer V cortical neurons, and 50–75 µm for DG GCs; the number of spines was scored in a segment of 25 µm for middle region cortical dendrites, and a segment of 20 µm for middle region DG GC dendrites. The distal regions were defined as 200–250 µm for layer V cortical neurons, and 100–125 µm for DG GCs; the number of spines was scored in a segment of 25 µm for distal cortical dendrites, and a segment of 20 µm for distal DG GC dendrites. A total of 100 layer V and DGs were scored for wild type, $Sema3F^{-/-}$ and $Npn-2^{-/-}$; four animals were used per genotype. A total of 160 layer V or DGs were scored for wild-type and $PlexA3^{-/-}$ mutants, and 155 layer V and 165 DG $PlexA4^{-/-}$ neurons; four animals were used per genotype. A total of 75 neurons (25 neurons from each of three animals) was scored in $Npn1^{Sema-}$ mutants and heterozygote controls. Secondary (or oblique) and basal dendritic *in vivo* spine densities are averages of at least two branches scored from a 25 µm segment of dendrite. All *in vitro* spine density quantification was obtained from a 25 µm dendrite segment, starting 25 µm from the cell soma and extending distally, either from the primary apical dendrite of cortical neurons or from one of the primary DG GC hippocampal neuron dendrites. For assessment of excitatory synaptic puncta in neurons maintained in primary cell culture, only PSD-95 puncta with directly adjacent vGlut1 puncta were scored as colocalized vGlut1/PSD-95 puncta. For inhibitory synapses, all gephyrin-positive puncta colocalized with vGAT puncta were scored on both the soma and dendritic processes of dentate gyrus neurons.

Sholl analysis. All Golgi-labelled image stacks were taken with a Zeiss Axioskop2 microscope using a ×20 objective. Each neuron was reconstructed using Adobe Photoshop CS3, and all analyses were performed using the ImageJ Sholl Analysis Plugin (http://www-biology.ucsd.edu/labs/ghosh/software/); the centre of all concentric circles defined as the centre of cell soma. For cortical basal dendrite

Sholl analysis, the starting radius was 12.5 μm and the ending radius was 100 μm from the centre; the interval between consecutive radii was 12.5 μm. For DG neuron Sholl analysis the starting radius was 15 μm and the ending radius was 150 μm from the centre; the interval between consecutive radii was 15 μm.

AP cell-binding assays and fluorescence quantification. Wild-type or $Npn\text{-}2^{-/-}$ mutant primary cortical neurons were taken at E14.5, grown 18 days *in vitro* on glass coverslips, lightly fixed with methanol for 5 min at $-20\,^{\circ}C$ and blocked in 100 mM TBS, 100 mM NaCl, 4 mM $MgCl_2$ (pH 7.4) with 20% fetal bovine serum at room temperature for 1 h. The cells were then incubated in 5 nM AP–Sema3F or AP–Sema3A probes at $4\,^{\circ}C$ overnight, washed in TBS the next morning, and fixed with 4% paraformaldehyde (pH 7.4) for 5 min at room temperature. After blocking with 5% normal goat serum for 1 h, cells were incubated with primary rabbit polyclonal antibodies directed against human placental AP (1:500; Serotec) and mouse Map2 (1:500; Sigma, also known as Mtap2) overnight at $4\,^{\circ}C$ and subsequently incubated with secondary fluorescent antibodies for 2 h at room temperature. The cells were washed in TBS several times before mounting with Vectashield hard-set fluorescence mounting medium (Vector Laboratories) and analysing with a Zeiss Axioskop2 Mot Plus, LSM 5 Pascal confocal microscope. We first quantified the fluorescence due to AP-ligand binding to apical and basal dendrites. Only neurons with apical and basal dendritic morphologies resembling pyramidal neurons were included in fluorescence quantification performed with ImageJ. The ratio of the fluorescence resulting from AP-ligand labelling of the apical dendrite over fluorescence resulting from labelling of all basal dendritic processes (AP-apical/AP-basal) was determined and divided by the ratio of fluorescence resulting from Map2 labelling of apical dendritic processes over labelling of all basal dendritic processes (Map2-apical/Map2-basal). If fluorescence intensities are equally distributed among all dendritic processes, then (AP-apical/AP-basal)/(Map2-apical/Map2-basal) is equal to 1. Next, we quantified the fluorescence intensity differences between primary apical versus secondary apical dendrites resulting from AP–Sema3F labelling. The fluorescence intensity resulting from AP–Sema3F labelling over anti-Map2 fluorescence of the primary apical dendrites (AP–Sema3F primary/Map2 primary) was compared to the fluorescence from AP–Sema3F labelling over anti-Map2 fluorescence of secondary apical dendrites (AP–Sema3F secondary/Map2 secondary). Measurements were obtained from a 10-μm segment of primary and secondary apical dendrites immediately adjacent to the branch point from the primary apical dendrite.

Electrophysiology. Mutant mice and their wild-type littermates aged 3–4 weeks were anaesthetized by isoflurane inhalation and decapitated. Brains were quickly dissected in ice-cold buffer containing 210.3 mM sucrose, 11 mM glucose, 2.5 mM KCl, 1 mM NaH_2PO_4, 26.2 mM $NaHCO_3$, 0.5 mM ascorbate, 2 mM *myo*-inositol, 0.5 mM $CaCl_2$ and 4 mM $MgCl_2$. Brains were vibratome-sectioned in the same solution at 350 μm and transferred to normal ACSF composed of 119 mM NaCl, 2.5 mM KCl, 1 mM NaH_2PO_4, 26.2 mM $NaHCO_3$, 11 mM glucose, 2 mM $CaCl_2$ and 2 mM $MgCl_2$. Slices were recovered at $35\,^{\circ}C$ for 1 h and then maintained at room temperature ($22\text{--}25\,^{\circ}C$). Neurons were targeted for whole-cell patch-clamp recording with borosilicate glass electrodes having a resistance of $3\text{--}6\,M\Omega$. The electrode internal solution was composed of 130 mM caesium methanesulphonate, 10 mM HEPES, 0.5 mM EGTA, 8 mM CsCl, 5 mM TEA-Cl, 1 mM QX-314, 10 mM Na phosphocreatine, 0.5 mM Na-GTP and 4 mM Na-ATP. Dentate gyrus granule cells were selected from the dorsomedial hippocampus and, where indicated, synaptic responses were evoked by stimulation of perforant pathway fibres with a matrix microelectrode. Cortical pyramidal neurons were selected from layer V of the posterior medial barrel subfield of primary somatosensory cortex and, where indicated, synaptic responses were evoked by stimulation of layer II/III inputs using a bipolar electrode to selectively examine paired pulse dynamics of apical dendritic synapses. During collection of AMPA receptor-mediated miniature EPSCs, external solution was supplemented with the following: 1 μM tetrodotoxin, 50 μM D,L-APV (2-amino-5-phosphonovalerate) and 100 μM picrotoxin. Data were acquired with a Multiclamp 700B and PCLAMP 10 software (Molecular Devices) at 10 kHz. Prior to mEPSC detection and analysis, current traces were low-pass filtered at 1 kHz. Events having amplitude of $2\times$ root mean square noise were detected using Mini Analysis (Synaptosoft). Kinetic measurements were performed on scaled, mean EPSC traces using a monoexponential decay function. Rise times correspond to 20–80% of peak amplitude. A total of 10 and 11 wild-type neurons and 12 and 10 neurons were recorded from layer V and dentate gyrus, respectively; $n = 2$ animals per genotype.

22. Saito, T. *In vivo* electroporation in the embryonic mouse central nervous system. *Nature Protocols* **1**, 1552–1558 (2006).
23. Fishell, G. & Hanashima, C. Pyramidal neurons grow up and change their mind. *Neuron* **57**, 333–338 (2008).
24. Inoue, T., Heo, W., Grimley, J., Wandless, T. & Meyer, T. An inducible translocation strategy to rapidly activate and inhibit small GTPase signaling pathways. *Nature Methods* **2**, 415–418 (2005).

LETTERS

Novel mutant-selective EGFR kinase inhibitors against EGFR T790M

Wenjun Zhou[1,2]*, Dalia Ercan[3,4]*, Liang Chen[3,4]*, Cai-Hong Yun[1,2]*, Danan Li[3,4], Marzia Capelletti[3,4],
Alexis B. Cortot[3,4], Lucian Chirieac[5], Roxana E. Iacob[6,7], Robert Padera[5], John R. Engen[6,7], Kwok-Kin Wong[3,4,8,9],
Michael J. Eck[1,2], Nathanael S. Gray[1,2] & Pasi A. Jänne[3,4,8]

The clinical efficacy of epidermal growth factor receptor (EGFR) kinase inhibitors in *EGFR*-mutant non-small-cell lung cancer (NSCLC) is limited by the development of drug-resistance mutations, including the gatekeeper T790M mutation[1–3]. Strategies targeting EGFR T790M with irreversible inhibitors have had limited success and are associated with toxicity due to concurrent inhibition of wild-type EGFR[4,5]. All current EGFR inhibitors possess a structurally related quinazoline-based core scaffold and were identified as ATP-competitive inhibitors of wild-type EGFR. Here we identify a covalent pyrimidine EGFR inhibitor by screening an irreversible kinase inhibitor library specifically against EGFR T790M. These agents are 30- to 100-fold more potent against EGFR T790M, and up to 100-fold less potent against wild-type EGFR, than quinazoline-based EGFR inhibitors *in vitro*. They are also effective in murine models of lung cancer driven by EGFR T790M. Co-crystallization studies reveal a structural basis for the increased potency and mutant selectivity of these agents. These mutant-selective irreversible EGFR kinase inhibitors may be clinically more effective and better tolerated than quinazoline-based inhibitors. Our findings demonstrate that functional pharmacological screens against clinically important mutant kinases represent a powerful strategy to identify new classes of mutant-selective kinase inhibitors.

EGFR kinase inhibitors, gefitinib and erlotinib, are effective clinical therapies for NSCLCs that harbour activating mutations in the *EGFR* kinase domain[1,6]. The most common *EGFR* mutations, L858R and delE746_A750, impart both an increased affinity for gefitinib or erlotinib and a decreased affinity for ATP relative to wild-type (WT) EGFR[7,8]. The clinical efficacy of gefitinib or erlotinib is, however, ultimately limited by the development of acquired drug resistance such as by mutation of the gatekeeper T790 residue (T790M), which is detected in 50% of clinically resistant patients[2,3]. Unlike the analogous T315I mutation in ABL, which introduces a steric impediment for imatinib binding, EGFR T790M only modestly affects gefitinib binding. However, more importantly, it restores the affinity for ATP, similar to that of WT EGFR[9].

Most EGFR inhibitors are based on a 4-anilinoquinazoline core scaffold and were initially identified as ATP-competitive inhibitors of WT EGFR. They include irreversible inhibitors that, unlike gefitinib, contain an electrophilic functionality that undergoes a Michael addition reaction with a conserved cysteine residue present in EGFR (Cys 797). The covalent nature of these compounds allows them to achieve greater occupancy of the ATP site relative to reversible

inhibitors, thus providing the ability to inhibit EGFR T790M in pre-clinical models, despite the increased ATP affinity conferred by this secondary mutation[4,10,11]. However, all current irreversible inhibitors are less potent in cell-line models harbouring *EGFR* T790M than those with an *EGFR*-activating mutation alone (Supplementary Fig. 1) and, at clinically achievable concentrations, these agents do not inhibit EGFR T790M *in vitro*[10–13]. Because the ATP affinity of EGFR T790M is similar to WT EGFR, the concentration of quinazoline-based EGFR inhibitors required to inhibit EGFR T790M will also effectively inhibit WT EGFR. In patients, this concurrent inhibition of WT EGFR results in skin rash and diarrhoea, and limits the ability to achieve plasma concentrations sufficient to inhibit EGFR T790M. Consequently, the clinical efficacy of the irreversible EGFR inhibitors CI-1033, HKI-272 and PF00299804 has been limited, especially in patients with gefitinib- or erlotinib-resistant NSCLC, and the dose-limiting toxicity has been diarrhoea and skin rash[5,14,15].

We hypothesized that the anilinoquinazoline scaffold may not be the most potent or specific for inhibiting EGFR T790M because it relies on the small size and hydrogen bonding interactions with the gatekeeper threonine of WT EGFR. We prepared a focused library of common kinase inhibitor core scaffolds where one of the side chains was modified with an acrylamide group at a position that molecular modelling predicted to react with Cys 797. This library was screened for compounds that could inhibit the growth of both gefitinib-resistant (PC9GR; delE746_A750/T790M) and -sensitive (PC9; delE746_A750) cell lines but were not toxic up to 10 μM against A549 (*KRAS* mutant) or H3122 (*EML4-ALK*) cells. We compared our findings with both reversible (gefitinib) and irreversible EGFR inhibitors (CL-387,785 and HKI-272). Three closely related pyrimidines, WZ3146, WZ4002 and WZ8040, were identified from the screen that possessed up to a 300-fold lower half-maximum inhibitory concentration (IC$_{50}$) against the PC9GR cells compared with clinical-stage inhibitors such as HKI-272 (Fig. 1a, b and Supplementary Table 1). We observed a similar increased potency of the WZ compounds in the H1975 (L858R/T790M) cell line and in Ba/F3 cells harbouring *EGFR* T790M (Fig. 1b and Supplementary Tables 1 and 2). The increased cellular potency correlated with inhibition of EGFR, AKT and ERK1/2 phosphorylation in NSCLC cell lines (Fig. 1c and Supplementary Fig. 2) and with the more potent inhibition of EGFR phosphorylation by WZ4002 in NIH-3T3 cells expressing different *EGFR* T790M mutant alleles (Fig. 1d and Supplementary Fig. 3). The profile against ERBB2 was markedly different: the WZ compounds were less potent than CL-387,785 or HKI-272 (Supplementary Tables 1 and 2) and did not inhibit ERBB2
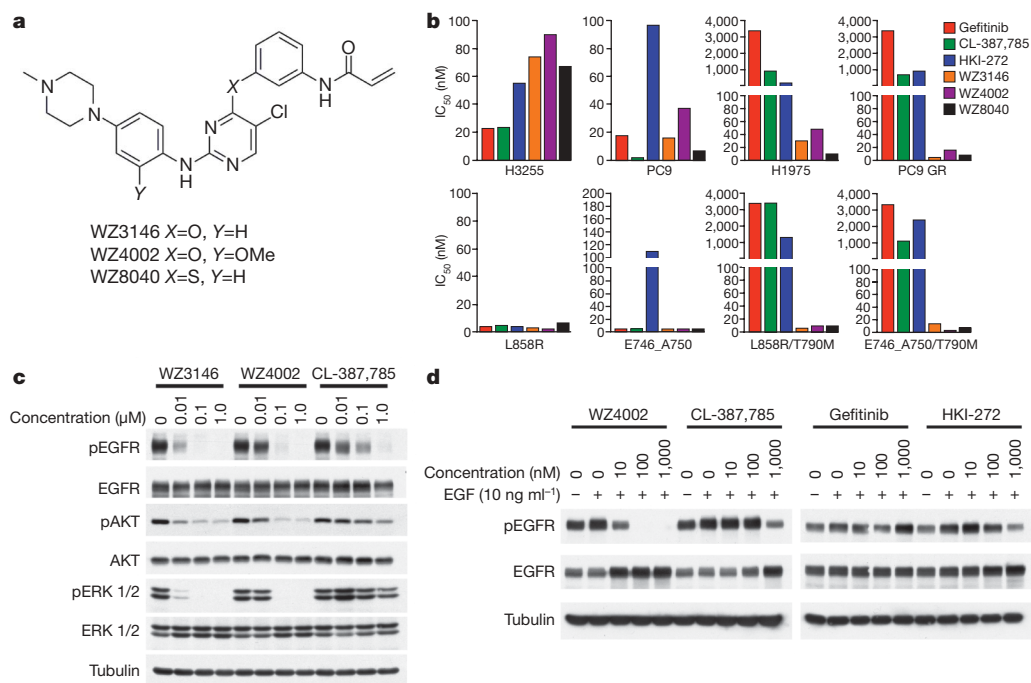
Figure 1 | WZ4002, WZ3146 and WZ8040 are novel EGFR inhibitors, suppress the growth of *EGFR*-T790M-containing cell lines and inhibit EGFR phosphorylation. a, Chemical structures of the WZ compounds. **b**, IC_{50} values for NSCLC cell lines (top) and Ba/F3 cells (bottom), with genotypes corresponding to the NSCLC cell lines, treated with indicated drugs. Growth was assessed using the MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) survival assay. **c**, Comparison of WZ3146, WZ4002 and CL-387,785 on EGFR signalling in

PC9 GR cells. The cells were treated with the indicated concentrations of each drug for 16 h. Cell extracts were immunoblotted to detect the indicated proteins. **d**, Comparison of EGFR inhibitors on EGFR phosphorylation in 3T3 cells expressing del E746_A750/T90M. The cells were treated with indicated concentrations of each drug for 16 h and stimulated with EGF (10 ng ml⁻¹) 15 min before lysis. Cell extracts were immunoblotted to detect the indicated proteins.

phosphorylation in 3T3 cells expressing the ERBB2 gatekeeper (T798I) mutation (data not shown). Analysis of recombinant EGFR T790M kinase incubated with WZ3146 by electrospray mass spectrometry revealed stoichiometric addition of one inhibitor molecule to the

protein. Analysis of a pepsin digest of the modified protein by tandem mass spectrometry identified Cys 797 as the site of modification, thus verifying covalent bond formation between WZ3146 and EGFR (Supplementary Fig. 4).

Figure 2 | WZ4002 is less potent than quinazoline EGFR inhibitors against WT EGFR *in vitro* and *in vivo*. a, EGFR vIII Ba/F3 cells treated with WZ or quinazoline EGFR inhibitors. The mean ($n = 6$) and standard deviation is plotted for each drug and concentration. **b**, Comparison of EGFR inhibitors on EGFR phosphorylation in 3T3 cells expressing WT EGFR. The cells were treated with indicated concentrations of each drug for 16 h and stimulated with EGF (10 ng ml⁻¹) 15 min before lysis. Cell extracts were

immunoblotted to detect the indicated proteins. **c**, Immunohistochemical analysis of hair bulb from erlotinib- or WZ4002-treated mice using EGFR and pY1173 EGFR. Only treatment with erlotinib results in significant inhibition of EGFR phosphorylation. Scale bar, 50 μm. **d**, Quantification of frequency of phospho-EGFR staining from vehicle- ($n = 3$), erlotinib- ($n = 3$) and WZ4002- ($n = 2$) treated mice. The means and standard deviations are plotted for drug treatment.

We profiled WZ3146 and WZ4002 against a panel of 400 kinases using the Ambit kinome screening platform (Supplementary Table 3 and Supplementary Fig. 5). For WZ4002, kinases that exhibited greater than 95% inhibition relative to the dimethylsulphoxide control (Ambit score <5) at 10 μM were selected for measurement of their dissociation constants (Supplementary Table 3). In addition to EGFR, we observed potent inhibition of several of the ten kinases that possess a cysteine at the same position as EGFR, including a subset of the TEC-family kinases (Supplementary Fig. 6). Cross-reactivity with BMX has been reported for irreversible quinazoline-derived EGFR inhibitors[16]. To confirm whether the observed binding activity translated into cellular inhibition, WZ4002 and WZ3146 were profiled against Ba/F3 cells transformed with TEL fusions of BMX, BLK, JAK2 and JAK3. WZ4002, which possesses an ortho-methoxy group at the C2-aniline substituent, is more selective for EGFR than WZ3146 (Supplementary Table 4).

We next determined whether the increased potency of the WZ compounds against mutant EGFR also applied to WT EGFR. We used *EGFR* WT HN11 cells[17] and Ba/F3 cells harbouring the *EGFR* vIII mutation, which contains a WT kinase domain (Fig. 2a and Supplementary Table 2). These compounds were 3- to 100-fold less potent, with WZ4002 being least potent, than CL-387,785 and HKI-272 at inhibiting the growth of the *EGFR* WT cells. Furthermore, WZ4002 was 100-fold less effective at inhibiting phosphorylation of WT EGFR than the quinazoline inhibitors (Fig. 2b). Similarly, WZ4002 inhibited EGFR kinase activity of recombinant L858R/T790M protein more potently than that of WT EGFR, whereas the opposite was observed with HKI-272 and gefitinib (Supplementary Fig. 7a, b).

To understand better the potency and relative selectivity for EGFR T790M over WT EGFR, we determined the crystal structure of WZ4002 in complex with EGFR T790M (Fig. 3a, b, Supplementary Fig. 8 and Supplementary Table 5). The compound binds within the ATP-binding cleft of the enzyme, forming the expected covalent bond with Cys 797. As expected based upon co-structures of related pyrimidine-derived inhibitors with CDK2 (ref. 18), JNK1 (ref. 19) and FAK[20], the anilinopyrimidine core of WZ4002 forms a bidentate hydrogen bonding interaction with the 'hinge' residue Met 793 (Fig. 3b). The chlorine substituent on the pyrimidine ring contacts the mutant gatekeeper residue, Met 790. The hydrophobicity conferred by this mutation likely contributes to the potency of these compounds against the T790M mutant. The aniline ring forms a hydrophobic interaction with the α-carbon of Gly 796 and its methoxy substituent extends towards Leu 792 and Pro 794 in the hinge region. The greater selectivity of the WZ4002 compound likely derives from the fact that both JAK3 and TEC-family kinases have a bulkier residue (tyrosine in JAK3, phenylalanine in TEC-family kinases) in the position of Leu 792, which would be expected to interfere sterically with the methoxy group in WZ4002. The reactive acrylamide moeity and the linking phenyl ring comprise the other arm of the inhibitor. The 'linker' phenyl ring lies roughly perpendicular to the pyrimidine core; this orientation juxtaposes the acrylamide with the thiol of Cys 797 for covalent bond formation (Fig. 3b).

We further determined whether WZ4002 is effective *in vivo* by using mouse lung cancer models harbouring either *EGFR* L858R/T790M or Del E746_A750/T790M. We chose WZ4002 for the *in vivo* studies because *in vitro* it was least potent against WT EGFR and other cysteine-containing kinases (Supplementary Tables 2 and 4) but was effective against EGFR T790M. A pharmacokinetic study was performed to determine the achievable plasma concentration (429 ng ml^{-1}), half-life (2.5 h) and the oral bioavailability (24%) of WZ4002 (Supplementary Tables 6–8). In a pharmacodynamic study, WZ4002 effectively inhibited EGFR, AKT and ERK1/2 phosphorylation (Fig. 4a), which was associated with a significant increase in TdT-mediated dUTP nick-end labelling (TUNEL)-positive and a significant decrease in Ki67-positive cells compared with mice treated by vehicle alone (Fig. 4b, c). To evaluate whether WZ4002 imparted a differential effect on WT EGFR *in vivo*, we evaluated EGFR phosphorylation in the hair bulb from mouse skin after treatment with either erlotinib or WZ4002
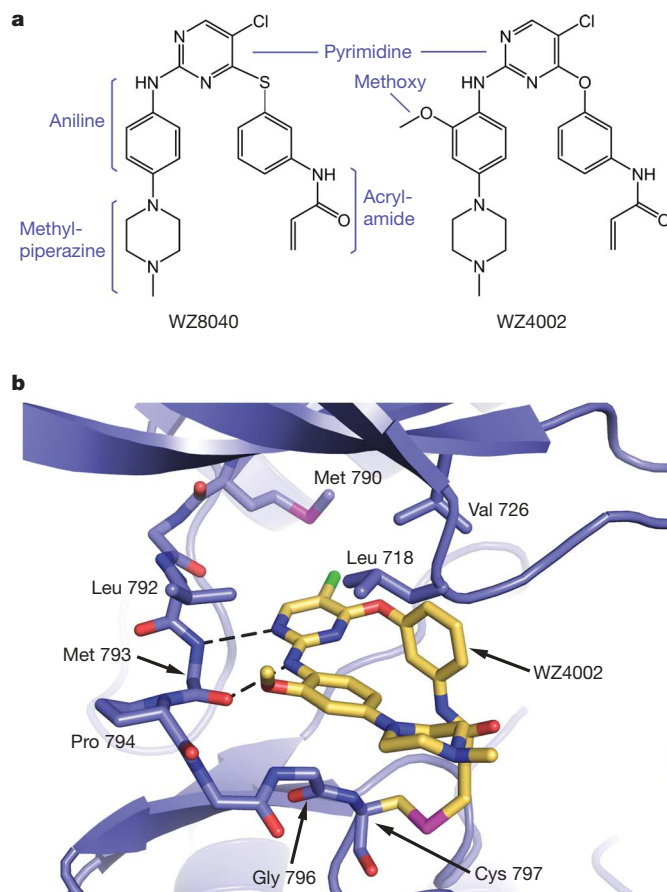

Figure 3 | Crystal structure of WZ4002 bound to EGFR T790M.
a, Chemical structures of WZ8040 and WZ4002 are shown schematically in a manner resembling the conformation adopted in complex with the kinase. **b**, Crystal structure of WZ4002 in complex with EGFR T790M mutant (PDB ID 3IKA). WZ4002 binds the active conformation of the kinase, with both the regulatory C-helix and the 'DFG' segment of the activation loop in their inward, active positions. The EGFR kinase is shown in a ribbon representation (blue) with the bound inhibitor in yellow. Side-chain and main-chain atoms are shown for selected residues that contact the compound. Expected hydrogen bonds to the backbone amide and carbonyl atoms of Met 793 are indicated by dashed lines. Note also the covalent bond with Cys 797. The structure was refined to a crystallographic R value of 21.3% (R_{free} = 25.4%) with data extending to 2.9-Å resolution (see Methods for further crystallographic details).

(Fig. 2c, d). Only erlotinib significantly inhibited EGFR phosphorylation in the hair bulb. In a 2-week efficacy study, WZ4002 treatment resulted in significant tumour regressions compared with vehicle alone in both T790M-containing murine models (Fig. 4d, e and Supplementary Fig. 9). Histological evaluation of the lungs after treatment confirmed significant resolution of the tumour nodules, with only a few small residual nodules and nodule remnants that had evidence of treatment effect with decreased cellularity and increased fibrosis consistent with remodelling/scarring (Fig. 4f). There were no signs of overt toxicity compared with mice treated by vehicle alone during the study as assessed by changes in weight (data not shown), serum creatine and total white blood cell count (Supplementary Fig. 10).

Our studies identify a novel structural class of EGFR kinase inhibitors that are effective *in vitro* and in *in vivo* models harbouring the *EGFR* T790M mutation. Given the marked activity in models with established *EGFR* T790M, we determined whether WZ4002 treatment could also prevent the development of *EGFR* T790M using *in vitro* models harbouring *EGFR*-activating mutations. Unlike with gefitinib or HKI-272, which when used at their achievable plasma concentrations lead to development of *EGFR* T790M *in vitro*[13,21,22],
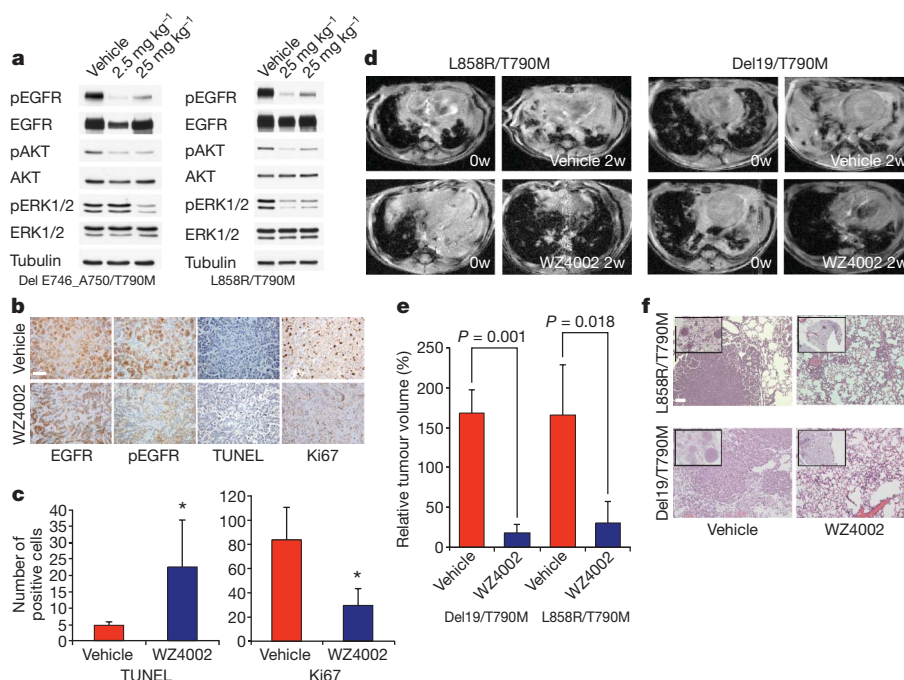
Figure 4 | WZ4002 inhibits EGFR phosphorylation and induces significant tumour regression in murine models of *EGFR* T790M. **a**, Two doses separated by 16 h of WZ4002 (2.5 mg kg^{-1} or 25 mg kg^{-1}) or vehicle were administered to EGFR delE746_A750/T790M or L858R/T790M mice with MRI-confirmed tumours. The mice were killed, the lungs isolated, grossly dissected and subjected to cell lysis. Cell extracts were immunoblotted to detect the indicated proteins. **b**, Immunohistochemical analyses of tumours from EGFR delE746_A750/T790M mice from **a** using indicated antibodies. Scale bar, 50 μm. **c**, Quantification of TUNEL- and Ki67-positive cells from tumour nodules ($n = 4$) from vehicle- and WZ4002-treated mice. The means

and standard deviations are plotted. *, $P < 0.05$. **d**, MRI images of vehicle- or WZ4002-treated mice at baseline (0 weeks: 0w) and after 2 weeks (2w) of treatment. **e**, Quantification of the relative tumour volume from MRI images from vehicle-treated mice (E746_A750/T790M ($n = 3$); L858R/T790M ($n = 4$)), and WZ4002-treated L858R/T790M ($n = 3$) and E746_A750/T790M ($n = 3$) mice. The means and standard deviations are plotted. **f**, Tumours from vehicle- and WZ4002-treated mice stained with haematoxylin and eosin. Low-power view (inset) demonstrates near-complete resolution of tumours in the WZ4002-treated mice. Scale bar, 100 μm.

we were unable to isolate any *EGFR*-T790M-containing clones from WZ4002-treated Ba/F3 or PC9 NSCLC cells (Supplementary Table 9). These findings suggest that WZ4002 could also be used as initial therapy for patients with *EGFR*-mutant NSCLC and may ultimately lead to a longer time to disease progression than currently achieved with gefitinib[1].

Our crystallographic studies provide insight into why WZ4002 is so much more effective against L858R/T790M than HKI-272. Although both share the irreversible component, the anilinopyrimi-dine scaffold of WZ4002 is an intrinsically better fit for the mutant gatekeeper methionine (Supplementary Fig. 11). To test this hypothesis further, we prepared WZ4003, a reversible analogue of WZ4002 that is non-reactive towards Cys 797. WZ4003 binds to the L858R/T790M mutant 100-fold more tightly than it does to the WT EGFR (Supplementary Table 10), confirming that the scaffold per se is indeed specific for the mutant kinase. Importantly, the WZ compounds rely on covalent bond formation for potent cellular inhibition, as evidenced by the 100-fold increase in the IC$_{50}$ of WZ4002 against the *EGFR* C797S mutants and by the significantly reduced cellular IC$_{50}$ of WZ4003 against T790M containing Ba/F3 cells (Supplementary Table 2). These observations highlight the importance of using a library of irreversible kinase inhibitors, as WZ4003 would not have been identified in the initial cellular screen.

Mutations, including those at the gatekeeper residue, are a common mechanism of drug resistance to kinase inhibitors. The current approach, using a cellular screen expressing the mutant kinase of interest, can be applied to identify novel agents specifically against drug resistance or oncogenic mutations implicated in human cancers. Such agents may truly be cancer selective, clinically more potent and less toxic than those that also concurrently inhibit the WT kinase. The agents described here are unique in that they inhibit both the drug sensitizing and resistance mutations but are selective against WT

EGFR[23–25]. Further studies are needed to determine whether this class of EGFR inhibitors will be clinically effective in patients with *EGFR*-mutant cancers harbouring *EGFR*-T790M-mediated acquired drug resistance.

METHODS SUMMARY

Kinase inhibitors. Gefitinib, CL-387,785 and HKI-272 were obtained from commercial sources. The WZ compounds were synthesized using a four-step chemical synthesis that is described in detail in the Supplementary Methods. The final products were verified by ^{1}H nuclear magnetic resonance and liquid chromatography–mass spectrometry.

Cell lines. *EGFR* wild-type and mutant NSCLC, Ba/F3 cells and NIH-3T3 cells were cultured as previously described[10]. The PC9GR4 cells were generated as previously described and verified to contain EGFR delE746_A750/T790M by direct sequencing[21]. Cell proliferation and growth assays were performed using a colorimetric assay as previously described[26]. Site-directed mutagenesis was performed using the Quick Change Site-Directed Mutagenesis kit (Stratagene) according to the manufacturer's instructions.

EGFR kinase assays. *In vitro* inhibitory enzyme kinetic assays using recombinant EGFR L858R/T790M and WT protein and were performed using the ATP/NADH coupled assay system in a 96-well format as previously described[7].

Crystal structure determination and refinement. The structure of WZ4002 in complex with EGFR T790M was determined as previously described[9].

Mouse studies. All studies involving mice were approved by the Dana-Farber Cancer Institute Animal Care and Use Committee. EGFR-TL (T790M/L858R) mice were generated as previously described[4]. EGFR exon 19 deletion-T790M (TD)-inducible bitransgenic mice were similarly generated and characterized. Mice were treated either with vehicle (10% 1-methyl-2-pyrrolidinone:90% PEG-300) alone or WZ4002 at 25 mg kg^{-1} gavage daily. This was followed by magnetic resonance imaging (MRI) scanning as previously described[4,27]. Histology, immunohistochemistry and immunoblotting analyses were performed according to standard protocols, as previously described[4,10].

Generation of drug-resistant cells. *N*-ethyl-*N*-nitrosourea mutagenesis was performed using *EGFR* L858R and DelE746_A750 Ba/F3 cells as previously described[28]. Treated cells were expanded in 100 nM WZ4002, 1 μM WZ4002,

200 nM HKI-272 or 1 μM gefitnib. Resistant clones were isolated and further characterized.

1. Mok, T. S. et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N. Engl. J. Med. 361, 947–957 (2009).
2. Pao, W. et al. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. PLoS Med. 2, 1–11 (2005).
3. Kobayashi, S. et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N. Engl. J. Med. 352, 786–792 (2005).
4. Li, D. et al. Bronchial and peripheral murine lung carcinomas induced by T790M–L858R mutant EGFR respond to HKI-272 and rapamycin combination therapy. Cancer Cell 12, 81–93 (2007).
5. Besse, B. et al. Neratinib (HKI-272), an irreversible pan-ErbB receptor tyrosine kinase inhibitor: preliminary results of a phase 2 trial in patients with advanced non-small cell lung cancer. Eur. J. Cancer Suppl. 6, 64, abstr. 203 (2008).
6. Rosell, R. et al. Screening for epidermal growth factor receptor mutations in lung cancer. N. Engl. J. Med. 361, 958–967 (2009).
7. Yun, C. H. et al. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell 11, 217–227 (2007).
8. Carey, K. D. et al. Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. Cancer Res. 66, 8163–8171 (2006).
9. Yun, C. H. et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc. Natl Acad. Sci. USA 105, 2070–2075 (2008).
10. Engelman, J. A. et al. PF00299804, an irreversible pan-ERBB inhibitor, is effective in lung cancer models with EGFR and ERBB2 mutations that are resistant to gefitinib. Cancer Res. 67, 11924–11932 (2007).
11. Li, D. et al. BIBW2992, an irreversible EGFR/HER2 inhibitor highly effective in preclinical lung cancer models. Oncogene 27, 4702–4711 (2008).
12. Yuza, Y. et al. Allele-dependent variation in the relative cellular potency of distinct EGFR inhibitors. Cancer Biol. Ther. 6, 661–667 (2007).
13. Godin-Heymann, N. et al. The T790M 'gatekeeper' mutation in EGFR mediates resistance to low concentrations of an irreversible EGFR inhibitor. Mol. Cancer Ther. 7, 874–879 (2008).
14. Janne, P. A. et al. Preliminary activity and safety results from a phase I clinical trial of PF-00299804, an irreversible pan-HER inhibitor, in patients (pts) with NSCLC. J. Clin. Oncol. 26 (Suppl.), abstr 8027 (2008).
15. Janne, P. A. et al. Multicenter, randomized, phase II trial of CI-1033, an irreversible pan-ERBB inhibitor, for previously treated advanced non small-cell lung cancer. J. Clin. Oncol. 25, 3936–3944 (2007).
16. Hur, W. et al. Clinical stage EGFR inhibitors irreversibly alkylate Bmx kinase. Bioorg. Med. Chem. Lett. 18, 5916–5919 (2008).
17. Yonesaka, K. et al. Autocrine production of amphiregulin predicts sensitivity to both gefitinib and cetuximab in EGFR wild-type cancers. Clin. Cancer Res. 14, 6963–6973 (2008).
18. Breault, G. A. et al. Cyclin-dependent kinase 4 inhibitors as a treatment for cancer. Part 2: identification and optimisation of substituted 2,4-bis anilino pyrimidines. Bioorg. Med. Chem. Lett. 13, 2961–2966 (2003).
19. Alam, M. et al. Synthesis and SAR of aminopyrimidines as novel c-Jun N-terminal kinase (JNK) inhibitors. Bioorg. Med. Chem. Lett. 17, 3463–3467 (2007).
20. Lietha, D. & Eck, M. J. Crystal structures of the FAK kinase in complex with TAE226 and related bis-anilino pyrimidine inhibitors reveal a helical DFG conformation. PLoS One 3, e3800 (2008).
21. Ogino, A. et al. Emergence of epidermal growth factor receptor T790M mutation during chronic exposure to gefitinib in a non small cell lung cancer cell line. Cancer Res. 67, 7807–7814 (2007).
22. Engelman, J. A. et al. Allelic dilution obscures detection of a biologically significant resistance mutation in EGFR-amplified lung cancer. J. Clin. Invest. 116, 2695–2706 (2006).
23. Tsai, J. et al. Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. Proc. Natl Acad. Sci. USA 105, 3041–3046 (2008).
24. O'Hare, T. et al. SGX393 inhibits the CML mutant Bcr-AblT315I and preempts in vitro resistance when combined with nilotinib or dasatinib. Proc. Natl Acad. Sci. USA 105, 5507–5512 (2008).
25. Gontarewicz, A. et al. PHA-680626 exhibits anti-proliferative and pro-apoptotic activity on imatinib-resistant chronic myeloid leukemia cell lines and primary CD34+ cells by inhibition of both Bcr-Abl tyrosine kinase and Aurora kinases. Leuk. Res. 32, 1857–1865 (2008).
26. Engelman, J. A. et al. MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science 316, 1039–1043 (2007).
27. Ji, H. et al. The impact of human EGFR kinase domain mutations on lung tumorigenesis and in vivo sensitivity to EGFR-targeted therapies. Cancer Cell 9, 485–495 (2006).
28. Bradeen, H. A. et al. Comparison of imatinib mesylate, dasatinib (BMS-354825), and nilotinib (AMN107) in an N-ethyl-N-nitrosourea (ENU)-based mutagenesis screen: high efficacy of drug combinations. Blood 108, 2332–2338 (2006).

LETTERS

DNA nanomechanics allows direct digital detection of complementary DNA and microRNA targets

Sudhir Husale[1], Henrik H. J. Persson[2] & Ozgur Sahin[1]

Techniques to detect and quantify DNA and RNA molecules in biological samples have had a central role in genomics research[1–3]. Over the past decade, several techniques have been developed to improve detection performance and reduce the cost of genetic analysis[4–10]. In particular, significant advances in label-free methods have been reported[11–17]. Yet detection of DNA molecules at concentrations below the femtomolar level requires amplified detection schemes[1,8]. Here we report a unique nanomechanical response of hybridized DNA and RNA molecules that serves as an intrinsic molecular label. Nanomechanical measurements on a microarray surface have sufficient background signal rejection to allow direct detection and counting of hybridized molecules. The digital response of the sensor provides a large dynamic range that is critical for gene expression profiling. We have measured differential expressions of microRNAs in tumour samples; such measurements have been shown to help discriminate between the tissue origins of metastatic tumours[18]. Two hundred picograms of total RNA is found to be sufficient for this analysis. In addition, the limit of detection in pure samples is found to be one attomolar. These results suggest that nanomechanical read-out of microarrays promises attomolar-level sensitivity and large dynamic range for the analysis of gene expression, while eliminating biochemical manipulations, amplification and labelling.

In a microarray experiment, single-stranded probe DNA molecules with known sequences are immobilized on a surface at pre-defined locations. Hybridization of probe and target molecules is routinely measured by fluorescence, but it can also be detected by exploiting various nanoscale phenomena[19], including nanomechanics. For example, changes in surface stress[12] and added mass of target molecules[13], electrical forces[20], topographical changes in self-assembled bar-coded nucleic-acid probe tiles[21] and hydration-induced surface tension[15] have been used for hybridization detection.

Our nanomechanical detection scheme relies on the changes in the stiffness of DNA molecules immobilized on a surface before and after hybridization. We used a recently developed atomic force microscope (AFM) technique[22,23] to investigate the stiffness of single- and double-stranded DNA molecules, as illustrated in Fig. 1a. The resulting stiffness maps, such as in Fig. 1b, show hybridized DNA molecules as dark-brown spots. This identification is verified by a series of experiments with non-complementary targets and complementary targets with varying concentrations.

When the sharp tip of the AFM cantilever presses against the molecules on the surface, interaction forces increase at a rate proportional to local stiffness. The two force curves plotted in Fig. 1c show that the surface is stiffer in the single-stranded region. The responses of molecules to external forces are determined not only by their intrinsic mechanical properties but also by their conformations on the surface. The higher stiffness observed in single-stranded DNA is

largely due to the influence of the gold substrate. Single-stranded molecules either lie flat on the substrate or are squeezed flat by the tip without offering much resistance, so the tip essentially feels the stiffness of the substrate. This hypothesis is supported by the height measurements, which show that hybridized molecules seem taller than single-stranded molecules (Supplementary Information). The greater height of hybridized molecules also results in spot sizes larger than the physical size of the molecules, owing to tip convolution. The changes in height between single-stranded and hybridized molecules have also been used to detect hybridization[24,25]; however, this required target concentrations on the nanomolar to micromolar level. Because height measurements are particularly sensitive to the roughness of the substrate surface, atomically flat substrates had to be used. Measurement of a local material property such as stiffness, however, provides the specificity necessary to discriminate between single hybridized molecules on the surface.

Figure 1d shows stiffness values calculated for the dashed profile in Fig. 1b. The stiffness values for single- and double-stranded molecules are well separated from each other. In addition, the values measured at different spots are sufficiently uniform to provide a clear signature of hybridized molecules. Computer analysis of the stiffness map using this signature generated a binary image in which hybridized target molecules can be counted (Fig. 1e). RNA–DNA hybrids exhibit the same mechanical signature as hybridized DNA (Supplementary Information).

We carried out hybridization experiments with varying target DNA concentrations and with two different areas of probe immobilization. The results plotted in Fig. 1f show the dependency of the surface density of hybridized molecules on target concentration and spot size; lower target concentrations and larger immobilization areas produce fewer hybridized molecules for a given scan area. Concentration levels from 1 nM to 1 aM are detectible. In addition, no hybridized molecules are found with the non-matching sequence. These levels represent enhancements of three to eight orders of magnitude in the detection limit and detection range, relative to the previously reported direct detection methods[11–16,20,21].

To investigate the performance of our technique against the complex background of a biological sample, we analysed total RNAs extracted from tumour tissues. Detection and quantification of RNA requires the use of large amounts of starting material or additional enzymatic steps involving reverse transcription and amplification. Furthermore, small regulatory RNA molecules such as microRNAs (miRNAs)[26] are not directly compatible with conventional amplification schemes[27]. Therefore, the approach we present here can simplify the analysis of miRNA content in biological samples to a great extent.

We measured differential expressions of two miRNAs, hsa-mir-205 (*Homo sapiens* miR-205) and hsa-mir-194 (*Homo sapiens* miR-194-1 or miR-194-2), in colon and bladder tumours by analysing

[1]The Rowland Institute at Harvard, Harvard University, Cambridge, Massachusetts 02142, USA. [2]Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA.
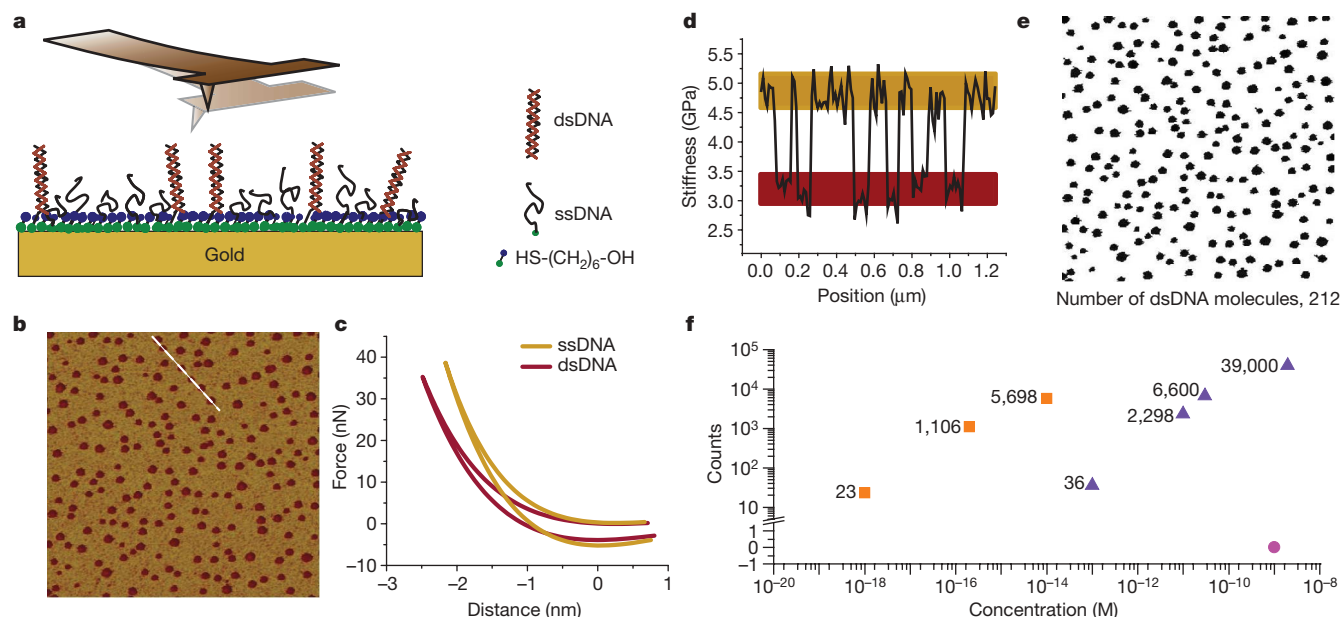
Figure 1 | Nanomechanical detection of DNA hybridization. a,b, Schematic of the experiment (**a**) and stiffness map (**b**). After hybridization, the surface of the array spot is scanned with the AFM to generate the stiffness map. Scan size, 3 μm. Hybridized molecules are measured to be less stiff and appear as dark brown spots. dsDNA, double-stranded DNA; ssDNA, single-stranded DNA. **c,** Stiffness at each pixel in **b** is calculated from force–distance curves. **d,** Effective stiffness values calculated for the dashed profile in **b. e,** Binary image computer-generated from stiffness values, in which the hybridized molecules can be counted. **f,** Numbers of hybridized molecules found in 10-μm-wide scan areas, plotted as a function of target concentration and for two probe immobilization areas (squares, 50 μm; triangles, 5 mm). Hybridization against a non-matching sequence is indicated by the pink circle.

total RNAs extracted from tissue samples. The extraction process includes disruption of cells and purification of RNAs. Material isolated in this way is routinely used for gene expression profiling. miRNAs exhibit expression patterns that can be used to predict the tissue origins of metastatic tumours[18], and the particular miRNAs we investigate here discriminate between tumours of gastrointestinal (colon) origin and those of non-gastrointestinal epithelial (bladder) origin. The numbers of RNA–DNA hybrids measured are plotted in Fig. 2a for hsa-mir-205 and in Fig. 2b for hsa-mir-194. (Numerical values from each measurement are provided in Supplementary Information.) The chart in Fig. 2c compares average expression levels normalized to 10 pg. The data show that hsa-mir-205 is expressed more in bladder tumour sample than in colon tumour sample, and that the opposite is true of hsa-mir-194. The observed inversion in differential expression in these miRNAs is in agreement with the

results of ref. 18. We further verified this result by conventional microarray analysis of the tumour samples (NCode Human miRNA Microarray V3, Invitrogen). The average numbers of miRNAs are also in broad agreement with the numbers derived from quantitative PCR experiments[27]. We note that the inversion in the differential expressions of the two miRNAs was detectable at amounts of total RNA as low as 0.2 ng. This represents a significant reduction in the amount of total RNA necessary relative to conventional microarray experiments, in which amounts in excess of 1 μg are typically required. These arrays are generally printed on glass slides and use spot sizes of around 100 μm × 100 μm. Their unprinted areas are blocked to eliminate non-specific adsorption.

An important advantage of our nanomechanical detection technique is its compatibility with multiplexed analysis in microarray platforms. Multiplexed analysis demands the generation of stiffness
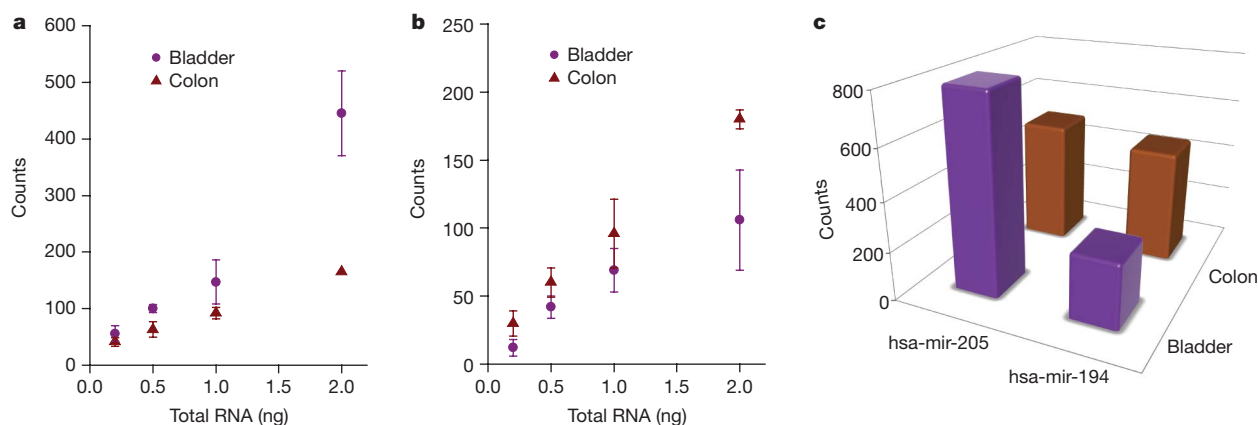

Figure 2 | Direct detection and quantification of miRNA expression in cancer tissues. Total RNAs isolated from bladder and colon tumour samples are analysed for miRNA content. **a,b,** Average numbers of target molecules found for varying amounts of total RNA are plotted for hsa-mir-205 (**a**) and hsa-mir-194 (**b**) (10-μm scan, 200-μm immobilization). Standard deviations (error bars) are calculated whenever two or more measurements are available from experiments carried out in triplicate. **c,** Comparison of expression levels in each sample, estimated by normalizing the data to 10 pg, averaging and scaling to entire area of immobilization.
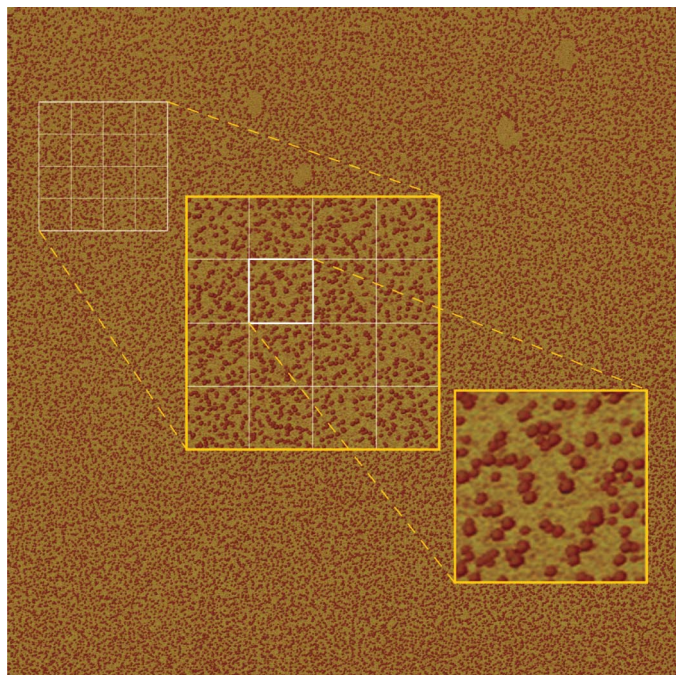
Figure 3 | High-throughput, robust nanomechanical read-out for multiplexed detection. $25 \times 25\ \mu m^2$ stiffness map at 6-nm pixel resolution, showing individual hybridized molecules. Scan time, 2 h. Target DNA concentration is 32 pM and immobilization area is $5 \times 5\ mm^2$. The number of double-stranded DNA molecules is ~41,250. The rectangular grid illustrates the possibility of sharing the scan area among different microarray spots. Insets show two levels of magnification (original stiffness data; not a rescan). The original image is given separately in Supplementary Information.

maps and the counting of hybridized molecules for each array spot. Our instrument is not optimized for scan speed, yet it provides the throughput, reliability and robustness for multiplexed analysis. As a demonstration, in Fig. 3 we show a 25-μm-wide stiffness map of hybridized molecules at 6-nm pixel size that was generated in 2 h. In principle, a scan area of this size is sufficient to read out an array with a thousand spots by sampling 800-nm-wide regions in each array spot. This level of multiplexing is sufficient for whole-genome miRNA expression profiling and for many clinical applications. Throughput can be further enhanced by combining our high-spatial-resolution read-out scheme with 'DNA nanoarrays' formed by dip-pen lithography[28] and supramolecular nanostamping[29]. Multiplexed analysis can also create more complex situations resulting from probes with secondary structures, mismatches, potential unusual base pairing and free termini. We have analysed several probe sequences that are representative of these cases and validated the utility of our stiffness-based hybridization-detection principle under these situations (Supplementary Information).

To conclude, we have seen that the nanomechanical signatures of hybridized DNA and RNA molecules have excellent background rejection, allowing digital read-out with an attomolar-level detection limit in micro- and nanoarray platforms while eliminating external labelling and enzymatic manipulations of the target molecules. Furthermore, the analysis of miRNA sampled from two tumour tissues illustrates the potential of this technique in medical applications. This combines the possibility of working with extremely small amounts of genetic material with substantial simplifications and cost reduction in genetic analysis.

METHODS SUMMARY

Sample preparation. We prepared the probe chips by immobilizing single-stranded and thiolated DNA molecules on a gold-coated silicon substrate. Immobilization areas were defined lithographically using a standard lift-off

procedure. After passivating the surface with mercaptohexanol[30], we immersed the chips in a solution containing target DNA or total RNA (with no labels or any other chemical modifications) for overnight hybridization, washed them with a buffer solution and dried them under a stream of nitrogen. Each sample was visually inspected and those that appeared contaminated or scratched were not included in the AFM analysis.

AFM analysis. We imaged samples in tapping-mode AFM under ambient conditions. T-shaped cantilevers were used to generate time-varying tip–sample forces. The effective elastic modulus was derived from these waveforms using our previously described mathematical procedure[22]. This procedure provides quantitative values for samples with flat surfaces. In the case of the tip interacting with DNA molecules, the calculated values provide a measure of stiffness, but they do not correspond to true elastic modulus values.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Higuchi, R., Dollinger, G., Walsh, P. S. & Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Nature Biotechnol.* **10**, 413–417 (1992).
2. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
3. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
4. Taton, T. A., Mirkin, C. A. & Letsinger, R. L. Scanometric DNA array detection with nanoparticle probes. *Science* **289**, 1757–1760 (2000).
5. Fritz, J., Cooper, E. B., Gaudet, S., Sorger, P. K. & Manalis, S. R. Electronic detection of DNA by its intrinsic molecular charge. *Proc. Natl Acad. Sci. USA* **99**, 14142–14146 (2002).
6. Cao, Y. C., Jin, R. & Mirkin, C. A. Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection. *Science* **297**, 1536–1540 (2002).
7. Park, S. J., Taton, T. A. & Mirkin, C. A. Array-based electrical detection of DNA with nanoparticle probes. *Science* **295**, 1503–1506 (2002).
8. Nam, J. M., Stoeva, S. I. & Mirkin, C. A. Bio-bar-code-based DNA detection with PCR-like sensitivity. *J. Am. Chem. Soc.* **126**, 5932–5933 (2004).
9. Fang, S., Lee, H. J., Wark, A. W. & Corn, R. M. Attomole microarray detection of microRNAs by nanoparticle-amplified SPR imaging measurements of surface polyadenylation reactions. *J. Am. Chem. Soc.* **128**, 14044–14046 (2006).
10. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnol.* **26**, 317–325 (2008).
11. Hahm, J.-i. & Lieber, C. M. Direct ultrasensitive electrical detection of DNA and DNA sequence variations using nanowire nanosensors. *Nano Lett.* **4**, 51–54 (2004).
12. Fritz, J. *et al.* Translating biomolecular recognition into nanomechanics. *Science* **288**, 316–318 (2000).
13. Ilic, B. *et al.* Enumeration of DNA molecules bound to a nanomechanical oscillator. *Nano Lett.* **5**, 925–929 (2005).
14. Zhang, J. *et al.* Rapid and label-free nanomechanical detection of biomarker transcripts in human RNA. *Nature Nanotechnol.* **1**, 214–220 (2006).
15. Mertens, J. *et al.* Label-free detection of DNA hybridization based on hydration-induced tension in nucleic acid films. *Nature Nanotechnol.* **3**, 301–307 (2008).
16. Clack, N. G., Salaita, K. & Groves, J. T. Electrostatic readout of DNA microarrays with charged microspheres. *Nature Biotechnol.* **26**, 825–830 (2008).
17. Burg, T. P. *et al.* Weighing of biomolecules, single cells and single nanoparticles in fluid. *Nature* **446**, 1066–1069 (2007).
18. Rosenfeld, N. *et al.* MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnol.* **26**, 462–469 (2008).
19. Cheng, M. M. C. *et al.* Nanotechnologies for biomolecular detection and medical diagnostics. *Curr. Opin. Chem. Biol.* **10**, 11–19 (2006).
20. Sinensky, A. K., & Belcher, A. M. Label-free and high-resolution protein/DNA nanoarray analysis using Kelvin probe force microscopy. *Nature Nanotechnol.* **2**, 653–659 (2007).
21. Ke, Y., Lindsay, S., Chang, Y., Liu, Y. & Yan, H. Self-assembled water-soluble nucleic acid probe tiles for label-free RNA hybridization assays. *Science* **319**, 180–183 (2008).
22. Sahin, O., Su, C., Magonov, S., Quate, C. F. & Solgaard, O. An atomic force microscope tip designed to measure time-varying nanomechanical forces. *Nature Nanotechnol.* **2**, 507–514 (2007).
23. Sahin, O. & Erina, N. High-resolution and large dynamic range nanomechanical mapping in tapping-mode atomic force microscopy. *Nanotechnology* **19**, 445717 (2008).
24. Zhou, D. J., Sinniah, K., Abell, C. & Rayment, T. Label-free detection of DNA hybridization at the nanoscale: a highly sensitive and selective approach using atomic-force microscopy. *Angew. Chem. Int. Ed.* **42**, 4934–4937 (2003).
25. Mirmomtaz, E. *et al.*, *Nano Lett.* **8**, 4134–4139 (2008).
26. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Rev. Genet.* **5**, 522–531 (2004).
27. Chen, C. *et al.* Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.* **33**, e179 (2005).

28. Demers, L. M. *et al.* Direct patterning of modified oligonucleotides on metals and insulators by dip-pen nanolithography. *Science* **296,** 1836–1838 (2002).

29. Yu, A. A. *et al.* Supramolecular nanostamping: using DNA as movable type. *Nano Lett.* **5,** 1061–1064 (2005).

30. Herne, T. M. & Tarlov, M. J. Characterization of DNA probes immobilized on gold surfaces. *J. Am. Chem. Soc.* **119,** 8916–8920 (1997).

Author Contributions S.H. performed the experiments, prepared gold substrates and developed experimental protocols; H.H.J.P. and O.S. contributed to the experiments. S.H. and H.H.J.P. designed and performed the surface chemistry, O.S. designed the cantilevers and wrote the stiffness calculation program, and H.H.J.P. and O.S. designed the biological assay. O.S. directed the research and wrote the paper; all authors discussed the results and commented on the manuscript.

METHODS

Substrate preparation. We prepared $5 \times 5\ mm^2$ gold-coated silicon chips by sputter coating (Au, 20 nm; Ti, 10 nm; AJA sputtering system, Center for Nanoscale Systems, Harvard University) on Si wafers (3 inch; Silicon Quest International). A diamond scribe was used for cutting wafers into $5 \times 5\ mm^2$ chips. Substrates with $50 \times 50\ \mu m^2$ and $200 \times 200\ \mu m^2$ gold spots were defined by photolithography followed by a standard lift-off procedure.

Immobilization of single-stranded DNA on the gold surface. We pretreated gold chips using a UV/Ozone ProCleaner (Bioforce Nanosciences) for ~30 min and washed them with nanopure water (deionized and filtered using an inorganic membrane filter with 20-nm pores). Thiolated single-stranded DNA molecules (concentration, 1 μM) were immobilized on the freshly prepared gold surface through thiol–gold covalent bonding, using high-salt buffer (1 M NaCl, 50 mM phosphate buffer, pH 7.2). Immobilization was done at room temperature (24 °C) under constant mixing; incubation time, ~15 h. The immobilized single-stranded-DNA chip surfaces were passivated by washing with water and subsequent incubation with mercaptohexanol (1 mM in nanopure water) for ~1 h.

Synthetic DNA samples. All oligonucleotides were purchased from MWG Biotech or from Integrated DNA Technologies and were used within a month. Thiolated probes were reduced according to manufacturers' suggested protocols.

DNA hybridization. Before hybridization, immobilized single-stranded-DNA chips were thoroughly washed with nanopure water. DNA–DNA hybridization reactions were carried out at room temperature under constant mixing in $1 \times$ SSC buffer (150 mM NaCl, 15 mM sodium citrate, pH 7.0). The volume of the hybridizing solution was 200 μl for $5 \times 5\ mm^2$ gold chips and 1 ml for 50-μm gold spots. Finally, hybridized chips were washed with low-salt buffer (25 mM NaCl, 10 mM Tris-HCl, pH 7.4) and gently dried in a weak stream of nitrogen.

Total RNA hybridization. Total RNAs extracted from human bladder and colon tumours were purchased from Stratagene and kept at −70 °C until used. Samples were diluted with hybridization buffer ($3 \times$ SSPE with Tween 20, pH 7.4; Teknova) to their final total RNA concentrations in Fig. 2. Hybridizations were carried out on a shaker for ~16 h at 42 °C. All microcentrifuge tubes, tweezers and gold-coated silicon chips were cleaned with RNaseZap (Ambion) before immobilization of DNA molecules, and washed extensively with DNase- and RNase-free water (MP Biomedicals).

AFM measurement conditions. Tapping-mode AFM was used for the measurements under ambient conditions. Torsional harmonic cantilevers (HarmoniX Probes, Veeco Instruments) were used (resonance frequency, ~60 KHz) to generate time-varying tip–sample forces. The driving amplitude was adjusted so that peak tapping forces during the scan were approximately 30–50 nN (peak tapping forces were different for single- and double-stranded molecules). The set-point amplitude was set to ~60 nm.

Calculation of elastic modulus. The numerical procedure to reconstruct the tip–sample force waveform from the raw deflection signals involved averaging vibration waveforms over several consecutive oscillation cycles to reduce noise, correcting the effect of the torsional frequency response, eliminating the cross-talk from large vertical signals and fitting the tip–sample force–distance curves to a contact mechanics model. Details of these steps and data acquisition hardware are described in ref. 22. Calculations were carried out in LABVIEW. Calibration of the time-varying force waveform (volts-to-newtons conversion) was performed following ref. 31. The effective elastic moduli of single- and double-stranded DNAs in Fig. 1 were calculated for a tip radius of 7 nm. Subsequent images were calibrated by using the single-stranded background in Fig. 1 as a reference.

31. Sahin, O. Harnessing bifurcations in tapping-mode atomic force microscopy to calibrate time-varying tip-sample force measurements. *Rev. Sci. Instrum.* **78**, 103707 (2007).

Rational design of a structural and functional nitric oxide reductase

Natasha Yeung[1], Ying-Wu Lin[1], Yi-Gui Gao[2], Xuan Zhao[1], Brandy S. Russell[1], Lanyu Lei[1], Kyle D. Miner[3], Howard Robinson[4] & Yi Lu[1,3]

Protein design provides a rigorous test of our knowledge about proteins and allows the creation of novel enzymes for biotechnological applications. Whereas progress has been made in designing proteins that mimic native proteins structurally[1–3], it is more difficult to design functional proteins[4–8]. In comparison to recent successes in designing non-metalloproteins[4,6,7,9,10], it is even more challenging to rationally design metalloproteins that reproduce both the structure and function of native metalloenzymes[5,8,11–20]. This is because protein metal-binding sites are much more varied than non-metal-containing sites, in terms of different metal ion oxidation states, preferred geometry and metal ion ligand donor sets. Because of their variability, it has been difficult to predict metal-binding site properties *in silico*, as many of the parameters, such as force fields, are ill-defined. Therefore, the successful design of a structural and functional metalloprotein would greatly advance the field of protein design and our understanding of enzymes. Here we report a successful, rational design of a structural and functional model of a metalloprotein, nitric oxide reductase (NOR), by introducing three histidines and one glutamate, predicted as ligands in the active site of NOR, into the distal pocket of myoglobin. A crystal structure of the designed protein confirms that the minimized computer model contains a haem/non-haem Fe_B centre that is remarkably similar to that in the crystal structure. This designed protein also exhibits NO reduction activity, and so models both the structure and function of NOR, offering insight that the active site glutamate is required for both iron binding and activity. These results show that structural and functional metalloproteins can be rationally designed *in silico*.

NOR is a metalloenzyme in the denitrification pathway of anaerobic bacteria that is responsible for the two-electron reduction of NO to N_2O (ref. 21). It is a key enzyme in the nitrogen cycle that is critical for all life. Furthermore, well-studied denitrification enzymes may provide structural and spectroscopic models for mammalian enzymes that produce and utilize NO in signal transduction pathways. Progress in understanding NOR has been hampered by difficulties in obtaining the enzyme in high yield and the lack of a three-dimensional structure. Peptide sequence alignments and structural threading have suggested that NORs are structurally homologous to subunit I of haem-copper oxidases (HCOs)[22]. All six His ligands to the active site of HCO (two to the low spin haem, one to the high spin haem, and three to Cu_B) are conserved in NOR at the same predicted positions in HCO. The major structural difference is the replacement of a copper-binding site (Cu_B) with a non-haem iron-binding site (Fe_B) in NOR[21,23]. Additionally, several conserved glutamates (not found in HCOs), have been identified near the catalytic haem/non-haem Fe_B site of NOR, and are predicted to have a role in iron binding and/or catalysis[24,25]. Synthetic models of NOR have been reported[21,26,27]. Despite these

progresses, computational design and three-dimensional (3D) structural confirmation of structural and functional models remain elusive, and it is often difficult to address long-range non-covalent interactions important for enzymatic functions using these models.

We chose to use sperm whale myoglobin (Mb) as a scaffold protein in which to engineer an active site with NO reduction activity because it is much easier to prepare and crystallize than both native proteins and synthetic models of NOR. Myoglobin and its mutants are also more amenable to biophysical characterization due to the presence of a single metal centre, free from interference by other chromophores. Although sperm whale Mb contains a haem centre, it lacks the Fe_B centre and NO reduction activity of NOR.

To design an Fe_B centre in sperm whale Mb, we examined its distal pocket for locations to introduce the putative non-haem metal-binding site, consisting of three histidines and one glutamate. The histidine residues were introduced first[14], as crystal structures of HCOs containing three conserved histidines are readily available (for example, cytochrome *c* oxidase). Based on overlays of the minimized structure of sperm whale Mb and cytochrome *c* oxidase, the distal His 64 in sperm whale Mb was chosen as one of the histidines, while Leu 29 and Phe 43 were mutated to histidines, to produce a non-haem metal-binding site. The V68E mutation to sperm whale Mb was chosen on the basis of its proximity and angle to the haem and the three His. The minimized computer model of the resultant triple mutant (sperm whale Mb L29H, F43H, V68E), named Fe_BMb, is shown in Fig. 1a. The Fe_B site (modelled as Zn(II)) is within bonding distance to Nε of all three His residues (2.13 Å to H29, 2.20 Å to H43 and 2.09 Å to H64) and both O atoms of E68 (2.11 Å and 1.96 Å), indicating that the proposed mutations would support iron binding in myoglobin, forming an Fe_B site. A comparison of the computer model overlaid with the crystal structure of native cytochrome *c* oxidase (Supplementary Fig. 1a) suggests structural similarity between the two centres.

To test designed Fe_BMb for iron binding, Fe^{2+} was titrated into deoxy (reduced haem iron) Fe_BMb that contains haem, but not any metal ions in the Fe_B site. The Soret band of deoxy Fe_BMb shifted slightly from 433 nm to 434 nm with Fe^{2+} addition. Concurrently, the peak at 557 nm was split into two shallow bands at 550 nm and 572 nm (Fig. 2A, a). This observation provided evidence that Fe_BMb is capable of incorporating iron into its designed metal-binding site, an important step towards modelling the Fe_B site of NOR. In contrast, titration of Fe^{2+} into either wild-type Mb (Fig. 2A, b) or a Mb derivative with the three His but no Glu (called $Fe_BMb(-Glu)$, Fig. 2A, c) resulted in no changes to the ultraviolet–visible spectra, indicating that both the histidines and particularly the glutamate, are required for Fe^{2+} binding.

Further evidence of Fe^{2+} binding to the designed Fe_B site comes from a high resolution crystal structure (1.72 Å) of Fe(II)-Fe_BMb
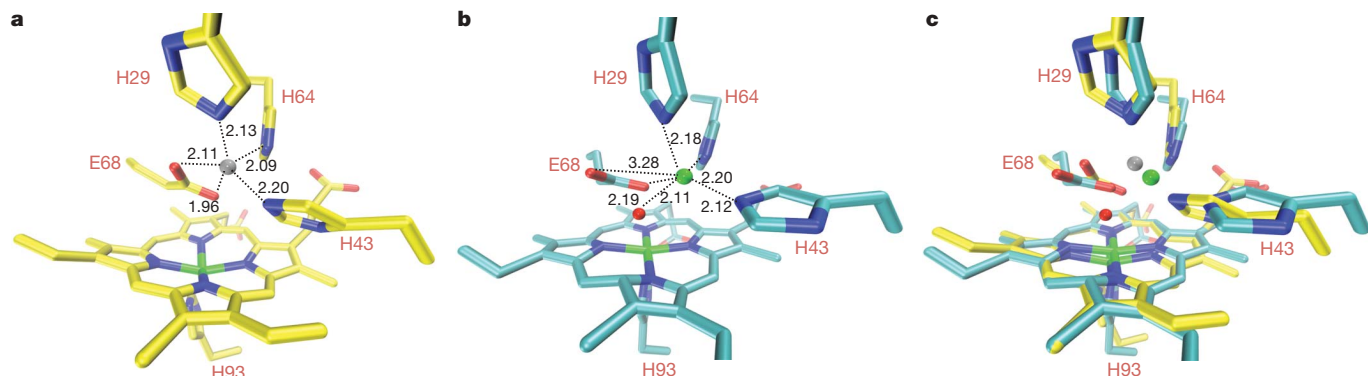

Figure 1 | Crystal structure of rationally designed Fe$_B$Mb overlays closely with minimized computer model. a, Minimized computer model of Fe$_B$Mb with Zn(II) in the Fe$_B$ site. **b**, Crystal structure of Fe(II)-Fe$_B$Mb collected at Fe-edge absorption (1.7309 Å) at the Brookhaven National Synchrotron Light Source X12C beamline (Upton, New York) with 1.72 Å resolution. The Fe···Fe distance is 4.82 Å, while the Fe–O–Fe angle is 115° (OE1 atom of E68). **c**, Overlay of Fe$_B$Mb model (yellow) with Fe(II)-Fe$_B$Mb crystal structure (cyan). In general, Fe(II) of the Fe$_B$ site is represented by a green sphere; Zn(II) (grey sphere) was used to model the Fe$_B$ site. A water molecule in the haem pocket is represented by a red sphere.

(Fig. 1b, Supplementary Figs 1 and 5 and Supplementary Table 1). The non-haem iron in the Fe$_B$ site is five coordinate, with bonds to all three histidines (2.18 Å to H29, 2.12 Å to H43 and 2.20 Å to H64), one O atom of E68 (2.19 Å) and one water molecule (2.11 Å). A weaker interaction between the non-haem iron and the other O atom of E68 at 3.28 Å is also present. The crystal structure presented here is consistent with the proposal that a glutamate in the active site of NOR helps stabilize iron binding to the Fe$_B$ site. In addition to structural roles, the glutamate, histidines and water in the active site may also provide the two protons required for NO reduction.

The overlay of the crystal structure of Fe(II)-Fe$_B$Mb with the computer model (Fig. 1c) is striking. Although His 43 does not overlay as well since it is located in a flexible loop, the other two His and Glu 68 overlay well with one another. Zn(II) in the model and Fe(II) in the Fe$_B$ site do not overlay exactly, but are in a similar position, 1.08 Å away from each other. Furthermore, the model did not predict the recruitment of a water molecule into the haem pocket, possibly because Zn(II) was used to model Fe(II), in the absence of Fe(II) parameters. Future development of Fe(II) parameters (not in a haem) will probably help resolve this issue. The haem in the computer model is puckered instead of planar, causing the iron to be pulled ~1 Å out of the haem plane. If out-of-plane distortion of the model is considered, the Fe$_B$ to haem iron distance is ~4.8 Å for both the model and crystal structure. Therefore, the distinct similarities between the crystal structure and computer model suggest that metalloprotein active sites can be rationally designed.

As the Fe$_B$ iron is only ~4.8 Å away from the haem iron, it is possible for spin-coupling to occur between the two irons, as is shown in native NOR[23,28]. To find out if the designed Fe$_B$Mb mimics this aspect of native NOR, we titrated different equivalents of Fe^{2+} into deoxy Fe$_B$Mb and then added blue copper azurin as a redox partner. In the absence of non-haem Fe^{2+} and azurin, deoxy Fe$_B$Mb displayed no EPR signal, as Fe(II) haem has an EPR-inactive integer electron spin in the perpendicular mode of EPR. Upon addition of Cu(II)-azurin, a
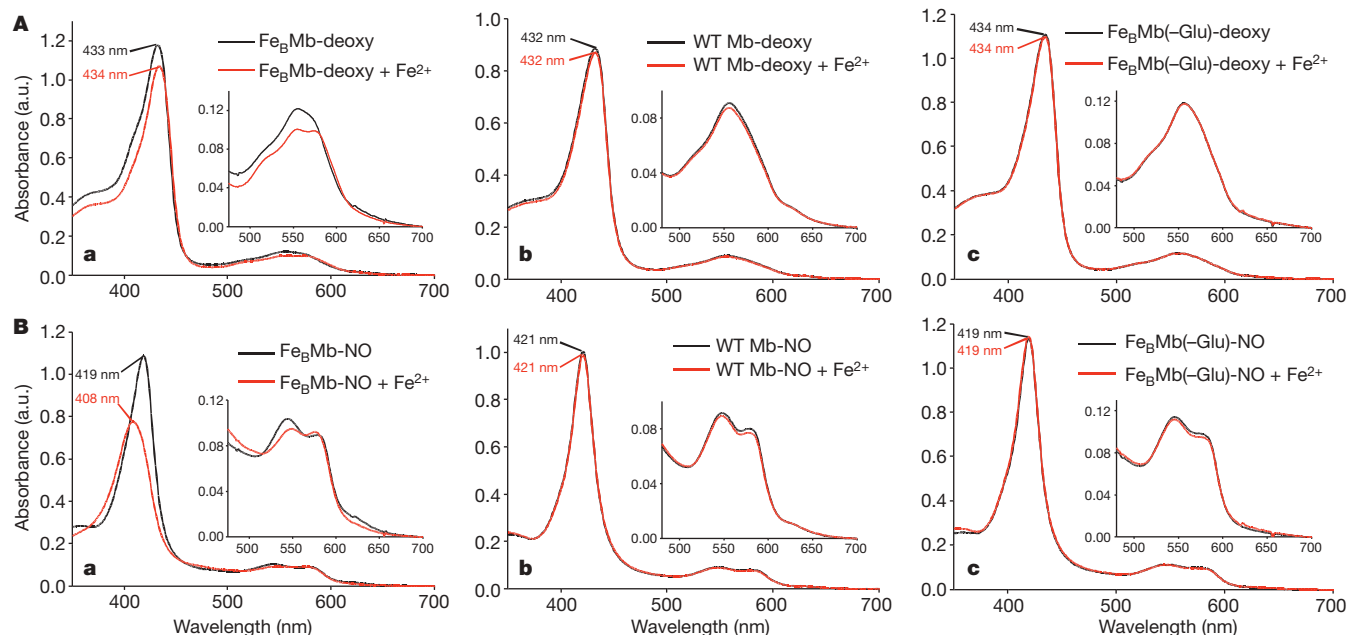

Figure 2 | Designed Fe$_B$Mb has NO reduction activity in the presence of Fe^{2+}. Control proteins show no evidence of NO activity under identical conditions. **A**, Ultraviolet-visible spectra of 10 μM deoxy Fe$_B$Mb (**a**), deoxy wild-type (WT) Mb (**b**), and deoxy Fe$_B$Mb(−Glu) (**c**) in the absence (black spectra) or presence (red spectra) of 2 equiv. Fe^{2+}. **B**, Ultraviolet-visible spectra of the NO reaction of 10 μM of deoxy Fe$_B$Mb (**a**), deoxy WT Mb (**b**), and deoxy Fe$_B$Mb(−Glu) (**c**) in the absence (black spectra) or presence (red spectra) of 2 equiv. Fe^{2+} after 20 min of incubation with ~17 equiv. NO. The deoxy protein was formed by reaction of met protein with excess dithionite in a glove box. Gel filtration purification was used to remove excess dithionite to preclude its reaction with NO.

high spin Fe(III) haem EPR signal centred at $g \approx 6$ was observed (Supplementary Fig. 2), suggesting that azurin can serve as a redox partner of Fe_BMb and oxidize the Fe(II) haem to Fe(III) haem. Interestingly, the intensity of the high spin haem signal decreased with Fe^{2+} binding to the Fe_B site of Fe_BMb, which suggests that Fe^{2+}, when oxidized by azurin, is spin-coupled to the Fe(III) haem, similar to that observed in native NOR. This conclusion is also supported by electrochemical redox potential determination, as the presence of iron in the Fe_B site resulted in a dramatic (112 mV) increase of the haem reduction potential (from -158 ± 4 mV to -46 ± 2 mV versus the normal hydrogen electrode, NHE; Supplementary Fig. 3).

To test its NO activity, deoxy Fe_BMb was reacted with excess NO (~ 17 equiv.) in the presence of Fe^{2+} and monitored by ultraviolet–visible spectroscopy (Fig. 2B). These conditions approximate one turnover as only two electrons are available for reaction, one from reduced haem iron and one from ferrous Fe_B. Furthermore, the reduction of NO to N_2O is highly thermodynamically favourable ($E_0' = +1.2$ V)[29], such that many common water soluble reductants can reduce NO non-enzymatically. Studying NO activity under single turnover conditions precludes non-enzymatic NO reduction and allows the reaction to be easily followed by spectroscopy. The appearance of visible peaks at 546 nm and 582 nm about 1 s after NO addition indicate fast NO binding. However, the NO-bound protein is not stable, and a species with peaks at 408 nm, 550 nm and 577 nm forms immediately after NO binding and is stable with no other changes in the spectrum over the course of 20 min (Fig. 2B, a, red spectrum). The Soret band at 408 nm is characteristic of Fe(III) haem and the visible peaks are characteristic of an NO-bound haem, similar to the NO-bound oxidized Fe(III) haem (met) form of Fe_BMb (Supplementary Fig. 4, black spectrum). Therefore, it appears that NO addition to Fe(II)-Fe_BMb results in a single species, namely the met-NO form. In contrast, when the Fe_B site of Fe_BMb is empty, NO addition results in NO binding with no additional changes in the ultraviolet–visible spectrum over 20 min (Fig. 2B, a, black spectrum). Therefore, these observations provide evidence that Fe_BMb with an intact Fe_B site can react with NO, whereas Fe_BMb with a vacant Fe_B site binds, but does not react with, NO.

As further controls, deoxy wild-type Mb lacking an Fe_B site, and an Fe_BMb derivative lacking the glutamate ($Fe_BMb(-Glu)$) were reacted with NO in the presence or absence of Fe^{2+}. After 20 min incubation with NO, the ultraviolet–visible spectra of both proteins in the presence of Fe^{2+} remains unchanged (Fig. 2B, b and c), suggesting the lack of NO activity for both proteins. These results are not surprising, as Fe^{2+} binding studies (Fig. 2A, b and c) indicate no evidence of iron binding for either protein. Therefore, the histidines and glutamate are important not only for iron binding, but also for NO reduction activity.

The product of NO reduction was identified by gas chromatography/mass spectrometry (GC/MS) under single turnover conditions by monitoring N_2O formation in the headspace of the solution. Because of the high solubility of N_2O (~ 25 mM in water at room temperature), this method cannot be used to quantify the rates of NO reduction under 1–2 turnover conditions (Supplementary Fig. 8). Typical MS of NO (molecular mass 30 Da) and N_2O (44 Da) are shown in Supplementary Fig. 6. When Fe(II)-Fe_BMb was incubated with NO, N_2O production was observed by the appearance of a second peak in the GC with a longer retention time than NO (Fig. 3a). Although the amount of N_2O formed by Fe(II)-Fe_BMb is low, it is consistently more than that produced by Fe^{2+} alone (without protein, Fig. 3b) over multiple samples. After no additional changes were observed over a period of 1–2 h, 2 equiv. of dithionite was added at 6 h to reduce the haem and Fe_B sites again, approximating a second turnover. Although considerable N_2O was produced in both samples (dithionite can non-enzymatically reduce NO to N_2O), Fe(II)-Fe_BMb produced ~ 2 times more N_2O than the control sample without protein, as observed in the chromatogram of only the molecular mass 44 Da component, representing the main ion peak of N_2O (Supplementary Fig. 7, bottom blue traces). As an additional control, wild-type Mb in the presence of Fe^{2+} was reacted with NO with no N_2O formation observed (Fig. 3c), indicating that non-specifically bound iron, if present, does not contribute to NO reduction activity. The yield of N_2O production by Fe(II)-Fe_BMb was estimated to be $\sim 30\%$. These results, together with ultraviolet–visible data (Fig. 2), indicate that designed Fe_BMb is a functional model of NOR.

We have rationally and successfully designed a protein model that both structurally and functionally mimics native NOR. We have demonstrated that both the glutamate and histidine ligands, which are conserved in NOR, are essential for iron binding and NO reduction activity. Being much smaller, easier to express in high yields, and free of
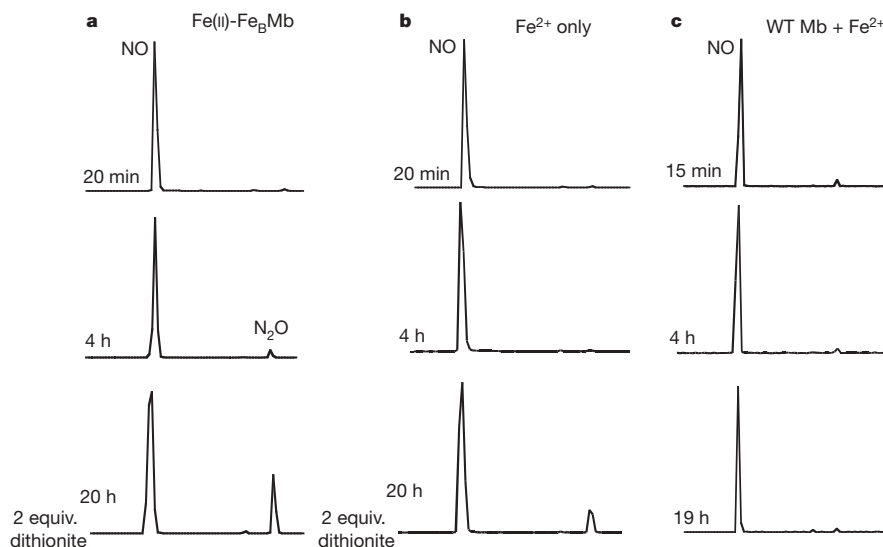

Figure 3 | Product of NO reaction with Fe_BMb is N_2O. a, Time dependent GC/MS measurements of N_2O formation by reaction of Fe(II)-Fe_BMb with NO. **b, c,** Control reactions of NO with Fe^{2+} only (**b**) and with Fe^{2+} and WT Mb (**c**). NO (~ 17 equiv.) was reacted with Fe(II)-Fe_BMb (0.6 mM protein, 1.5 mM or 2.5 equiv. Fe^{2+}), Fe^{2+} (1.5 mM, no protein), or with WT Mb (1 mM) and Fe^{2+} (2 mM or 2 equiv.). At 6 h, 2 equiv. dithionite (1.2 mM) was added to the Fe(II)-Fe_BMb and Fe^{2+} only samples and allowed to react to simulate a second turnover. NO_2 (molecular mass 46 Da) was not detected. GC peaks have been normalized. N_2O yield (30%) was determined from a comparison of the ratio of the NO:N_2O peaks of the molecular mass 30 Da:44 Da single ion chromatograms 2 h after dithionite addition (after which additional N_2O formation was not observed), to that of known dithionite concentrations (2 h after addition). Background N_2O formation (that is, Fe^{2+} without protein) was subtracted from that of Fe(II)-Fe_BMb to estimate the yield.

other chromophores, this designed protein will serve as an excellent model for mechanistic studies of NOR.

METHODS SUMMARY

Fe_BMb was purified from inclusion bodies by resolublization with 7 M $GuHCl$, reconstitution with a haem cyanide solution and dialysis against potassium phosphate, pH 7 (Methods). Protein mass was confirmed by SDS–PAGE and electrospray ionization mass spectrometry (ESI-MS). Ultraviolet visible spectra were collected at room temperature in buffered solution. EPR spectra were recorded as a glass at 4 K. Protein reduction potentials were measured by a spectroelectrochemical method as described previously[30]. Fe(II)-bound Fe_BMb crystals were grown anaerobically by the hanging drop method in an inert atmosphere glove box at room temperature. To test NO reduction activity, a solution of Fe_BMb was de-oxygenated and transferred to a glove box where it was reduced with dithionite. Excess reductant was removed with a size exclusion column (PD-10). $FeCl_2$ was added to reconstitute the Fe_B site. NO gas ($\sim$17 equiv.) was added to the resulting Fe(II)-Fe_BMb protein and allowed to react with stirring (anaerobically). N_2O production was monitored via GC/MS.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Regan, L. & DeGrado, W. F. Characterization of a helical protein designed from first principles. *Science* **241**, 976–978 (1988).
2. Hecht, M. H., Richardson, J. S., Richardson, D. C. & Ogden, R. C. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science* **249**, 884–891 (1990).
3. Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
4. Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl Acad. Sci. USA* **98**, 14274–14279 (2001).
5. Kaplan, J. & DeGrado, W. F. De novo design of catalytic proteins. *Proc. Natl Acad. Sci. USA* **101**, 11566–11570 (2004).
6. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
7. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
8. Lu, Y., Yeung, N., Sieracki, N. & Marshall, N. M. Design of functional metalloproteins. *Nature* **460**, 855–862 (2009).
9. Shifman, J. M. & Mayo, S. L. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc. Natl Acad. Sci. USA* **100**, 13274–13279 (2003).
10. Kang, S.-g. & Saven, J. G. Computational protein design: structure, function and combinatorial diversity. *Curr. Opin. Chem. Biol.* **11**, 329–334 (2007).
11. Robertson, D. E. *et al.* Design and synthesis of multi-haem proteins. *Nature* **368**, 425–432 (1994).
12. Hay, M., Richards, J. H. & Lu, Y. Construction and characterization of an azurin analog for the purple copper site in cytochrome *c* oxidase. *Proc. Natl Acad. Sci. USA* **93**, 461–464 (1996).
13. Yeung, B. K., Wang, X., Sigman, J. A., Petillo, P. A. & Lu, Y. Construction and characterization of a manganese-binding site in cytochrome *c* peroxidase: towards a novel manganese peroxidase. *Chem. Biol.* **4**, 215–221 (1997).
14. Sigman, J. A., Kwok, B. C. & Lu, Y. From myoglobin to heme-copper oxidase: design and engineering of a Cu_B center into sperm whale myoglobin. *J. Am. Chem. Soc.* **122**, 8192–8196 (2000).
15. Case, M. A. & McLendon, G. L. Metal-assembled modular proteins: toward functional protein design. *Acc. Chem. Res.* **37**, 754–762 (2004).
16. Cochran, F. V. *et al.* Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. *J. Am. Chem. Soc.* **127**, 1346–1347 (2005).
17. Watanabe, Y. & Hayashi, T. Functionalization of myoglobin. *Prog. Inorg. Chem.* **54**, 449–493 (2005).
18. Ghosh, D. & Pecoraro, V. L. Probing metal-protein interactions using a de novo design approach. *Curr. Opin. Chem. Biol.* **9**, 97–103 (2005).
19. Petros, A. K., Reddi, A. R., Kennedy, M. L., Hyslop, A. G. & Gibney, B. R. Femtomolar Zn(II) affinity in a peptide-based ligand designed to model thiolate-rich metalloprotein active sites. *Inorg. Chem.* **45**, 9941–9958 (2006).
20. Koder, R. L. *et al.* Design and engineering of an O_2 transport protein. *Nature* **458**, 305–309 (2009).
21. Wasser, I. M., de Vries, S., Moënne-Loccoz, P., Schröder, I. & Karlin, K. D. Nitric oxide in biological denitrification: Fe/Cu metalloenzyme and metal complex NO_x redox chemistry. *Chem. Rev.* **102**, 1201–1234 (2002).
22. van der Oost, J. *et al.* The heme-copper oxidase family consists of three distinct types of terminal oxidases and is related to nitric oxide reductase. *FEMS Microbiol. Lett.* **121**, 1–9 (1994).
23. Girsch, P. & de Vries, S. Purification and initial kinetic and spectroscopic characterization of NO reductase from *Paracoccus denitrificans*. *Biochim. Biophys. Acta* **1318**, 202–216 (1997).
24. Hendriks, J., Gohlke, U. & Saraste, M. From NO to OO: nitric oxide and dioxygen in bacterial respiration. *J. Bioenerg. Biomembr.* **30**, 15–24 (1998).
25. Watmough, N. J. *et al.* Nitric oxide in bacteria: synthesis and consumption. *Biochim. Biophys. Acta* **1411**, 456–474 (1999).
26. Wasser, I. M., Huang, H.-w., Moënne-Loccoz, P. & Karlin, K. D. Heme/non-heme diiron(II) complexes and O_2, CO, and NO adducts as reduced and substrate-bound models for the active site of bacterial nitric oxide reductase. *J. Am. Chem. Soc.* **127**, 3310–3320 (2005).
27. Collman, J. P. *et al.* A functional nitric oxide reductase model. *Proc. Natl Acad. Sci. USA* **105**, 15660–15665 (2008).
28. Moënne-Loccoz, P. *et al.* Nitric oxide reductase from *Paracoccus denitrificans* contains an oxo-bridged heme/non-heme diiron center. *J. Am. Chem. Soc.* **122**, 9344–9345 (2000).
29. Zumft, W. G. Nitric oxide reductases of prokaryotes with emphasis on the respiratory, heme-copper oxidase type. *J. Inorg. Biochem.* **99**, 194–215 (2005).
30. Zhao, X., Yeung, N., Wang, Z., Guo, Z. & Lu, Y. Effects of metal ions in the Cu_B center on the redox properties of heme in heme-copper oxidases: spectroelectrochemical studies of an engineered heme-copper center in myoglobin. *Biochemistry* **44**, 1210–1214 (2005).

Author Contributions N.Y. and Y.-W.L. performed most of the experimentation and wrote most of the manuscript. B.S.R. helped with the initial design of mutants and experiments. X.Z. and L.L. assisted in experimentation. K.D.M. performed computational modelling. Y.-G.G. guided crystallization and refined the crystal structure. H.R. collected crystal diffraction data. Y.L. designed, guided the project and edited the manuscript.

Author Information Crystallographic data for Fe_BMb have been deposited in the Protein Data Bank with accession number 3K9Z. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.L. (yi-lu@illinois.edu).

METHODS

Fe$_B$Mb computer modelling. A computer model of the designed protein Fe$_B$Mb was constructed with the L29H, F43H and V68E mutations to the crystal structure of sperm whale Mb (PDB: 1JP6) using the modelling extension in VMD (Visual Molecular Dynamics)[31]. Rendering was also performed in VMD. H43 was rotated about Cα to ensure interaction with the haem pocket as opposed to the bulk solvent. To assess the ability of the designed Fe$_B$ site to bind an iron ion, molecular dynamics simulations were performed. To mimic an iron ion, the water molecule found in the crystal structure of the distal pocket of sperm whale Mb was changed to a Zn(II) ion. The resulting computer model was then minimized with NAMD (Molecular Dynamics Simulator)[32] using 5,000 minimization steps at 0 K, then 10,000 molecular dynamics steps (1 fs per step) via an NVT ensemble (where the number of particles *N*, the volume *V*, and the temperature *T* of the system are kept constant) at 310 K.

Construction, expression and purification of Fe$_B$Mb. The triple mutant L29H/F43H/H64V/V68E of sperm whale myoglobin (called Fe$_B$Mb) was constructed using a protocol described previously[14,33] and confirmed by DNA sequencing at the Biotechnology Center of the University of Illinois. The DNA was transformed into BL-21(DE3) cells, expressed overnight, and harvested by centrifugation. The bacterial pellet from the equivalent of a 1 l growth was resuspended in 100 ml of 10 mM Tris (2-amino-2-hydroxymethyl-propane-1,3-diol) pH 8 containing 1 mM EDTA, 0.5 mM DTT (dithiothreitol), 35 U ml^{-1} DNaseI, 4 U ml^{-1} RNaseI, 200 mg lysozyme and a few crystals of PMSF (phenylmethylsulphonyl fluoride). The resultant solution was stirred at 4 °C overnight, diluted to 200 ml with the resuspension buffer, and centrifuged at 10,000g for 45 min. The resulting grey/white pellet was resuspended in 50 ml of 7 M guanidine HCl (GuHCl) and shaken at 4 °C for ~1 h. A solution of 25 mg KCN/7.5 mg haemin in 5 ml H$_2$O was prepared and diluted by rapid mixing into 450 ml of cold H$_2$O. Haem incorporation was achieved by dropwise addition of the protein/GuHCl to the haemin-CN solution over 30 min with rapid stirring at 4 °C. The resulting cloudy red solution was stirred for ~1 h after the final addition and centrifuged for 30 min at 8,000g to remove precipitate. The supernatant (500 ml) was dialysed against 8 l of H$_2$O and 4 l of 100 mM (ionic strength) potassium phosphate at pH 7 to remove excess GuHCl. The protein was refolded with three additional buffer changes with at least 6 h between changes. The protein was then centrifuged for 30 min at 8,000g and concentrated with a stirred cell concentrator (8000 MWCO membrane; Millipore) to 10 ml. The concentrated protein was then loaded onto a gel filtration column (Sephacryl S-100 High Resolution; GE Healthcare) equilibrated with 100 mM potassium phosphate pH 7 and eluted with the same buffer. The fractions with a high ($\geq$4.5) R/Z ratio (defined as an absorption ratio of A_{406nm}/A_{280nm}; the higher the R/Z ratio, the more pure the protein is) were pooled, concentrated, and dialysed twice against 100 mM potassium phosphate pH 7 containing 30 mM EDTA, then twice with 100 mM potassium phosphate pH 7 treated with a chelating resin (Chelex 100, iminodiacetic acid, Sigma) containing 100 mM NaCl, and finally three times with 100 mM potassium phosphate pH 7 treated with chelating resin to remove adventitious metal. The purity and identity of purified Fe$_B$Mb was confirmed by SDS–PAGE and ESI-MS (calculated molecular mass, 17,375 Da; experimental molecular mass, 17,376 ± 1 Da). Yield, ~13 mg protein per l of BL-21(DE3) bacterial growth.

UV-vis characterization of Fe$_B$Mb under one turnover conditions. Ultraviolet-visible spectra were collected on either a Cary 3E (Varian) spectrometer connected to a circulating water bath equipped with a digital temperature controller (Polyscience), or a Cary 5000 (Varian) spectrometer equipped with a Quantum Northwest Peltier temperature controller and stirring module. Kinetic ultraviolet-visible data were also collected to probe the reaction of NO with Fe$_B$Mb in the absence or presence of Fe^{2+} using a Hewlett-Packard 8453 spectrometer with the kinetics module. A protein solution was degassed and transferred to a glove box where it was reduced, purified by size exclusion column, and reconstituted with 2 molar equiv. of Fe^{2+} (FeCl$_2$, Alfa Aesar) when required.

EPR characterization of spin-coupling of Fe$_B$Mb in the presence of Fe^{2+}. To prepare EPR samples, Fe$_B$Mb was first reduced by dithionite in a glove box to form deoxy Fe$_B$Mb. Excess dithionite was removed with a size exclusion column. Different molar equivalents of Fe^{2+} were then titrated (that is, 0, 0.5, 1.0, 1.5, 2.0 equiv.) into the deoxy Fe$_B$Mb solution. After stirring for 20 min, 3 molar equivalents of oxidized WT Cu(II)-azurin was added to each sample to oxidize Fe$_B$Mb as well as the Fe(II) (Supplementary Fig. 2). After 5 min, the sample was then transferred into an EPR tube and flash frozen in liquid N$_2$. EPR samples contained 0.5 mM Fe$_B$Mb, 1.5 mM azurin and 6.8% glycerol in 50 mM Bis-Tris (bis(2-hydroxyethyl)iminotris(hydroxy-methyl)methane), pH 7.0. Control experiments of wild-type sperm whale Mb only, Fe$_B$Mb(−Glu) only and Fe^{2+} only were prepared under identical conditions. Spectra were recorded at 4 K on a Bruker ESP 300 equipped with an Oxford liquid helium cryostat and an ITC4 temperature controller.

Spectroelectrochemical determination of reduction potential. Protein reduction potentials were measured using an optically transparent thin layer electrode (OTTLE) as previously described[30,34], with the following changes. A 200 μM protein solution, ~0.6 ml, containing 2 molar equivalents of Fe^{2+} when required, and 100 mM NaCl in 50 mM Bis-Tris pH 7 buffer or 100 mM potassium phosphate pH 7 buffer without added NaCl were used. Each protein solution also contained 80 μM each of phenazine methosulphate (PMS, Sigma) and anthraquinone-2-sulphonate (AQS, Sigma) as electron transfer mediators. For the spectroelectro-chemistry study of deoxy Fe$_B$Mb in the presence of Fe^{2+} (starting from the reduced deoxy metal bound form and transitioning to the oxidized met form), the protein was reduced in the glove box with dithionite and purified of excess dithionite before mediator (AQS only) and Fe^{2+} addition, and subsequent transfer to the degassed OTTLE cell. The headspace of the OTTLE cell was purged with N$_2$ during the experiment to prevent O$_2$ contamination.

The potential of the working electrode was applied in the negative direction for deoxy Fe$_B$Mb and in the positive direction for deoxy Fe$_B$Mb + Fe^{2+}. The ultraviolet–visible spectrum of the protein at each potential was recorded with a Cary 3E or Cary 5000 spectrophotometer until no further spectral changes were observed.

Data analysis. The spectra from 300–800 nm were analysed by global analysis using singular value decomposition (SVD) and non-linear regression modelling software (SpecFit/32, Spectrum Software Associates). The reduction potential of each protein solution was obtained by fitting the data to the best fit model, an A ↔ B model, in SpecFit/32 and checked by generating a Nernst plot at a single wavelength[30]. The Ag/AgCl (3 M KCl) reduction potentials were converted to NHE potentials by adding 0.209 V. Since the reduction potentials obtained from the Nernst plots and from SpecFit/32 modelling were generally within ±16 mV of each other, only the potentials derived from SpecFit/32 were reported (Supplementary Fig. 3).

Purification of NO and NO saturated solution preparation. NO gas (Matheson Trigas, 99%) was purged of N$_2$O$_3$, NO$_2$ and higher N$_x$O$_y$ impurities by passing it through a bubbler containing 1 M NaOH and a second bubbler of water before it entered a reaction vessel filled with ultrapure water[35]. The reaction solution was first purged with helium to remove O$_2$ to make a ~1.1 mM saturated NO solution, as measured by spectroscopic titrations of a saturated solution of NO to oxy WT Mb to form met WT Mb. NO was added via a gas-tight syringe to a protein and/or metal solution.

X-ray crystallographic studies. Fe^{2+} incorporated Fe$_B$Mb was prepared as above. Fe$_B$Mb crystals were grown anaerobically in a glove box at room temperature. Red crystals were observed after 2–3 days. Diffractable crystals were obtained from a 1.5 mM protein solution in 20 mM potassium phosphate pH 7 mixed 1:1 with well buffer (0.2 M sodium acetate trihydrate, 0.1 M sodium cacodylate trihydrate pH 6.5 and 30% w/v PEG 8000). The well of the crystal tray was filled with 300 μl of the same solution. Fe(II)-Fe$_B$Mb crystals were soaked in a cryoprotectant solution and flash frozen in liquid N$_2$ in a glove box before data collection.

Data collection. Diffraction data of Fe(II)-Fe$_B$Mb were collected using Fe-edge absorption (1.7309 Å) as wavelengths of data collection at the Brookhaven National Synchrotron Light Source X12C beamline. All data were integrated using the program HKL2000 (ref. 36).

Crystal structure determination. The crystal structure was solved by the molecular replacement method using MOLREP in the CCP4 Package[37]. Refinement was performed using X-plor[38] and SHELX'97 (ref. 39). The position of Fe^{2+} in the Fe$_B$ site, and the E68 and H43 residues were rebuilt by the program O[40].

NO activity studies by GC/MS. NO reduction was monitored by GC/MS using an HP-Plot Q column (Aligent), installed in a Hewlett Packard 5890 gas chromatography system equipped with an online mass spectrometer (Micromass 70-VSE). As dithionite can reduce NO non-enzymatically, the NO activity of Fe$_B$Mb was probed under one turnover conditions whereby the two electrons required for reduction, one from the reduced Fe(II) haem and one from the ferrous Fe$_B$ site of Fe(II)-Fe$_B$Mb, were rate limiting. Where applicable, an extra 2 molar equivalents of dithionite was added to the protein/metal solution to induce an additional turnover. Glucose, glucose oxidase and catalase were added to scavenge residual O$_2$ as described[41]. Under anaerobic conditions, a 2.67 ml reaction sample consisting of 0.6 mM protein, 10 mM glucose, glucose oxidase (100 μg ml^{-1}) and catalase (50 μg ml^{-1}) in 20 mM Bis-Tris (treated with chelating agent) pH 7 was prepared in a Schlenk flask fitted with a septum. The head-space of the flask was then exchanged to helium. Purified NO gas (1 atm) was injected into the head space of the reaction flask via a gas-tight syringe and the reaction mixture was stirred at room temperature (the molar ratio of NO:protein ≈ 17:1). Typically, 250 μl of head space gas was analysed via GC/MS at different time intervals to observe NO and N$_2$O levels. MS of NO (molecular mass 30 Da) and N$_2$O (molecular mass 44 Da) are shown in Supplementary Fig. 6. NO$_2$ (molecular mass 46 Da) formation was not observed.

31. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* 14, 33–38 (1996).
32. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802 (2005).

33. Zhao, X., Yeung, N., Russell, B. S., Garner, D. K. & Lu, Y. Catalytic reduction of NO to N$_2$O by a designed heme-copper center in myoglobin: implications for the role of metal ions. *J. Am. Chem. Soc.* **128,** 6766–6767(2006).

34. Taboy, C. H., Bonaventura, C. & Crumbliss, A. L. Anaerobic oxidations of myoglobin and hemoglobin by spectroelectrochemistry. *Methods Enzymol.* **353,** 187–209 (2002).

35. Bonner, F. T. Nitric oxide gas. *Methods Enzymol.* **268,** 50–57 (1996).

36. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).

37. Vagin, A. & Teplyakov, A. *MOLREP*: an automated program for molecular replacement. *J. Appl. Crystallogr.* **30,** 1022–1025 (1997).

38. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).

39. Sheldrick, G. M. & Schneider, T. R. SHELXL: high-resolution refinement. *Methods Enzymol.* **277,** 319–343 (1997).

40. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).

41. Giuffrè, A. *et al.* The heme-copper oxidases of *Thermus thermophilus* catalyze the reduction of nitric oxide: evolutionary implications. *Proc. Natl Acad. Sci. USA* **96,** 14718–14723 (1999).

Q&A

Martin Cole takes over as chief of food and nutritional sciences at the Commonwealth Scientific and Industrial Research Organisation in Australia in January 2010.

How did you get started in the food business?

I am a microbiologist by training. My PhD, funded by the food company Colman's of Norwich, UK, was focused on the prevention of yeast spoilage. That joint project between the University of East Anglia and the Institute of Food Research in Norwich gave me my first taste of the complexity involved in making food, and the interplay between academia and industry. I realized I enjoyed bridging disciplines as diverse as food engineering, chemistry and microbiology.

What early career move had a lasting impact?

Early on, I joined the International Commission on Microbiological Specifications for Foods. It was a productive, yet small, group of about 20 microbiologists from all over the world. We developed crucial standards, sampling plans and methods for food safety, as well as new approaches to risk management. It was a great opportunity to learn from the top scientists in my field. I tell students it is important to get involved in your field's professional associations because they provide a unique opportunity to learn from peers.

What have your industry experiences taught you?

I think the most important things I learned from the private sector are leadership and management skills. But industry was also a good scientific training ground. My first boss at Unilever was an expert on bacterial spores and food safety. I've used those industry-specific scientific skills to help develop new approaches to risk management. For example, as chairman of the International Commission on Microbiological Specifications for Foods, I helped to introduce the concept of a universal food-safety objective, which translates a food-safety goal into a quantifiable maximum hazard level. It's a metric used to communicate food-safety levels between governments and industry.

How is the field of food science changing?

We're finding ways to bridge the gap between food science and nutritional science. We need to close the loop between food processing and understanding the effect food has in the body. To do that, we need to harness the molecular-science advances that will allow us to ultimately do some clinical work. At end of the day, we've got to see what effect food has in the human population. I currently work at the US National Center for Food Safety and Technology, a collaborative venture between the US Food and Drug Administration and industry. There, we are engaged in efforts to change how we look at food — beyond food safety to include the health and nutritional aspects of food.

What is the biggest challenge in food production?

The global demand for food is going to double by 2050, with processed food making up much of the trade in food. Creating sustainable ways to feed 9 billion people living mostly in cities in a responsible, sustainable way — all while climate gets more unpredictable — is the biggest challenge we face.

Can we sustainably produce enough food to feed 9 billion people?

We've certainly gone through the sorts of change needed to realize big gains in production in the past, specifically from 1960 to 2000. But I don't think we're going to see the next big jump in production without strong science and technology advances, in molecular genetics and crop production for example. When I lecture, I tell people it is a good time to get into food science if you want to make a difference in the world.

What research opportunities are there for food scientists?

As the food supply chain gets more complicated, so does the need to protect the health of the consumer. We need to develop ways to characterize and respond quickly to new hazards. We also need to reduce the waste from food spoilage, while determining how to preserve the most beneficial aspects of food. To that end, we need to understand the molecular mechanisms that underpin the nutritional aspects of food. The big challenge is doing that in a way that helps the general population make good food choices.

What types of collaborations do you plan to foster at the Commonwealth Scientific and Industrial Research Organisation (CSIRO)?

We need to partner with industry because, ultimately, we aren't going to produce healthy, safe food without industry doing it. At the the CSIRO, we will aim to translate scientific principles into things that industry can adopt. Young minds are also needed to do this research, so we'll focus on collaborating with university partners to develop schemes that bring young scientists into the field. ∎

Interview by Virginia Gewin.

IN BRIEF

Innovative energy

The US Advanced Research Projects Agency–Energy (ARPA-E) of the Department of Energy is seeking up to 20 energy-technology investigators for a two-year fellowship programme, which may lead to permanent positions at the department. Responsibilities will include seeking new areas for development and commercialization and helping to create energy policy. Inspired by the US Defense Advanced Research Projects Agency, ARPA-E supports high-risk 'transformational' research unlikely to be funded through traditional means.

Lab expansion opens

The Van Andel Institute, a non-profit biomedical research and educational organization in Grand Rapids, Michigan, has just opened a $178-million expansion — eight storeys of new lab space that will be fully operational by March 2010. In the next decade, as the institute expands from cancer research to translational medicine, administrators expect to add more than 20 principal-investigator posts, creating some 550 research and support-staff positions. Neurodegenerative disease, bone research, structural biology and pharmacogenomics are among the areas to be targeted.

J.DYKEHOUSE

Pharma job losses ranked

Of the dozens of drug companies worldwide that downsized this year, Pfizer has the dubious distinction of having made the most lay-offs. According to a top ten list compiled by newsletter *FiercePharma* in Washington DC, Pfizer slashed 19,500 jobs, partly as a consequence of its US$68-billion acquisition of Wyeth. Merck, which purchased Schering-Plough for $41 billion, came second with 16,000 lay-offs. Cost-cutting prompted Johnson & Johnson, in third place, and AstraZeneca, in fourth, to shed 8,900 and 7,400 jobs, respectively. GlaxoSmithKline cut 6,000 and Eli Lilly was sixth with 5,500.

Divine diseases

An act of faith.

Ian Watson

If God made man in His own image, what does that say about God's own anatomy?

When God made the world in 4004 BC, He was a young god, His rapidly inflating cosmic playground newly budded from the divine multiverse. Yet already He was wisely conscious of health issues and of a need for medical insurance. Ailments, of course, could not kill Him; divine diseases would be similar to the case of Prometheus and his pecked liver, which rapidly regenerated.
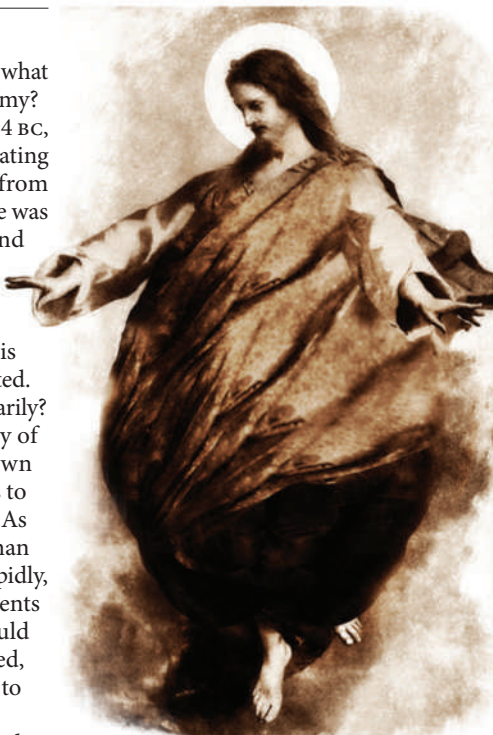
Yet why should He suffer unnecessarily?

Health care would come courtesy of His humans who would (in His own image) duly display all the ailments to which He might become vulnerable. As a result of their short lifespans, human beings would express much more rapidly, not to mention repeatedly, the ailments that might later assail Him, and would in due course, as civilization advanced, develop medical science with which to combat those ailments.

Dyspepsia and juvenile boils and acne afflicted the adolescent God badly during the days of the ancient Israelites, resulting in plagues and anger and storms and afflictions visited upon the antique, although simultaneously early, world. But presently Greeks such as Hippocrates and Theophrastus pioneered medical science, rooted in a notion of humours, which wasn't so far off course as God could be choleric, spleenful or melancholic (He never slept) and bilious (all those black holes in His cosmos, like stomach ulcers; all those luridly red and purple nebulae, like diseased organs of gas). Of course He could also be blessedly phlegmatic or sanguine (yet sanguine sometimes meant bloody-minded — blame not God, but rather the quirks of His physiology).

Consequently Greek diagnoses and cures, and those of the Arabs thereafter, were fairly sensible, although lacking sophistication and technology, which was what God awaited.

During a phlegmatic and benevolent era, God gave rise to His Son in a human body to bestow divine love upon His creation. However, in the long run this proved to be a serious mistake …

The problem wasn't that His Son was ungratefully crucified, to be resurrected and taken up bodily into Heaven; He would come again when the time was ripe. However, kicking off with the

Council of Ephesus in AD 431, and then proclaimed by the Fourth Lateran Council in 1215 under Pope Lunaticus, and fully ratified officially in the sixteenth century at the Council of Trent, the Catholic Church (and the unorthodox Orthodox churches too, in their own fashions) ordained that the bread and wine consumed in communion were miraculously transubstantiated into genuine flesh and blood.

Faith may move mountains, especially a little at a time. The weight of an individual wafer was only a few grams, but a million wafers per day would in due course add up to metric tonnes. In Heaven, alongside His Father, the Son began putting on weight as the transubstantiated bread (and subsequently wafers) became added to His body mass. Likewise the blood, to sustain the supplementary tissue. Be it noted that the newly created flesh must needs be transferred to the Son in Heaven after ingestion but before digestion; otherwise the divine flesh must pass onward into the bowels and become what normal nutrition normally becomes; which would be sacrilegious.

This weight gain didn't matter too much to begin with, as the Son had been fairly gaunt, just as most Palestinians were at the time of His birth. Not to mention that He was fairly tall (think El Greco). He could

benefit by some additional bodily tissue.

However, fast-forward a few hundred years more to the time when the numbers of the human race began to climb almost exponentially — half a billion, two billion, six billion. Many of these souls were occasional or even daily transubstantiators. (The Protestant Reformation might in this light be viewed as a slimming campaign, as Protestantism denied that bread and wafers actually changed; therefore they did not do so.)

Despite pruning of population by wars and, more effectively, by plagues such as the otherwise inexplicable Spanish influenza pandemic in the wake of the First World War, the Son's body mass in Heaven was increasing alarmingly. From overweight, to positively obese.

Yet God delayed the Second Coming because *only now* was medical science really getting into its stride. Penicillin and subsequent pharmaceuticals, radiotherapy, open-heart surgery, transplants, DNA diagnoses and all the other wonders of medical science: just wait another 50 years, and all of God's increasing bodily woes could be cured by His clever creatures who were both the laboratory rats for the ailments and the researchers into cures.

Wait a few more years. Yet because of transubstantiation, the obesity problem of the Son was becoming dire.

Finally the Second Coming was triggered prematurely by sheer geometrical and gravitational stress. As a result of the accumulated bulk of the Son, amounting by now to hundreds of metric tonnes, He couldn't reappear anywhere on the surface of Earth in a dignified way — Him with his vastly bloated body, out of which protruded a normal-sized head and hands and feet — nor perhaps without succumbing to an immediate heart attack.

Accordingly, it was in near-Earth orbit that the Second Coming occurred, a shining Jesus moonlet with its own miraculous warm atmosphere, rivalling Luna much farther away.

And the messages began: *Stop Transubstantiating! Too Much Mass!* ■

Ian Watson's most recent book *The Beloved of My Beloved* (NewCon Press, 2009), written with Italian surrealist Roberto Quaglia, may well be the first full-length genre fiction work by two authors with different mother tongues. Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a